Data Science

для кариериста

Жаклин Нолис
Эмили Робинсон

MANNING

Build a Career in Data Science

EMILY ROBINSON
AND JACQUELINE NOLIS

MANNING
SHELTER ISLAND

Жаклин Нолис
Эмили Робинсон

Data Science
для кариериста

ПИТЕР®

Санкт-Петербург • Москва • Минск

2021

ББК 32.973.233.02
УДК 004.62
Н80

Нолис Жаклин, Робинсон Эмили

Н80 Data Science для карьериста. — СПб.: Питер, 2021. — 368 с.: ил. — (Серия «Библиотека программиста»).

ISBN 978-5-4461-1734-5

Все мы хотим построить успешную карьеру. Как найти ключ к долгосрочному успеху в Data Science? Для этого понадобятся не только технические ноу-хау, но и правильные «мягкие навыки». Лишь объединив оба этих компонента, можно стать востребованным специалистом.Узнайте, как получить первую работу в Data Science и превратиться в ценного сотрудника высокого уровня! Четкие и простые инструкции научат вас составлять потрясающие резюме и легко проходить самые сложные интервью. Data Science стремительно меняется, поэтому поддерживать стабильную работу проектов, адаптировать их к потребностям компании и работать со сложными стейкхолдерами не так уж и легко. Опытные дата-сайентисты делятся идеями, которые помогут реализовать ваши ожидания, справиться с неудачами и спланировать карьерный путь.

16+ (В соответствии с Федеральным законом от 29 декабря 2010 г. № 436-ФЗ.)

ББК 32.973.233.02
УДК 004.62

ISBN 978-1617296246 англ.
ISBN 978-5-4461-1734-5

Оглавление

От Эмили для Майкла
и
от Жаклин для Хизер, Эмбер и Лауры
за любовь и поддержку, которую вы давали нам
на всем этом пути

Предисловие

«Как мне устроиться на такую же работу, как у вас?»

Нам как опытным дата-сайентистам постоянно задают этот вопрос. Порой он звучит прямо, а в других случаях нас спрашивают о том, какие решения мы принимали в течение карьерного пути, чтобы оказаться на этом месте. На самом деле люди, задающие подобные вопросы, постоянно испытывают трудности, так как ресурсов, объясняющих, как встать на путь Data Science или расти профессионально в этом направлении, очень мало. Многие дата-сайентисты ищут помощь по вопросам карьеры, но зачастую не находят внятных ответов.

Хотя в блогах мы постили тактические советы о том, что делать в определенные моменты работы в Data Science (DS), мы также решили разобраться с отсутствием адекватного текста, описывающего весь карьерный путь в этой области от начала до конца. Эта книга призвана помочь тысячам людей, которые слышат о Data Science и о машинном обучении, но не знают, с чего начать, а также тем, кто уже занят в этой области и хочет понять, как продвинуться по карьерной лестнице.

Мы были рады возможности поучаствовать в создании этой книги. Нам обеим казалось, что наш опыт и точки зрения дополняли друг друга и помогли в написании лучшей книги для вас. Мы — это:

- *Жаклин Нолис* (Jacqueline Nolis). Я получила степень бакалавра и магистра математических наук, а также кандидатскую степень в области исследования операций. Когда я начинала работать, такого понятия, как *Data Science* (*DS*), еще не было, и мне пришлось выстраивать свой карьерный путь одновременно с попытками определения этой области. Теперь я работаю консультантом и помогаю компаниям растить команды, занимающиеся DS.
- *Эмили Робинсон* (Emily Robinson). Я получила степень бакалавра в области теории принятия решений и степень магистра менеджмента. Окончив трехмесячный курс по Data Science в 2016 году, я начала работать в этой сфере, специализируясь на A/B-тестировании. Сейчас я работаю старшим дата-сайтентистом в компании Warby Parker и занимаюсь некоторыми проектами компании.

На своем карьерном пути мы создавали портфолио проектов и испытывали стресс от адаптации на новой работе. Когда нас не брали на желаемую должность, нам было обидно. Когда наш анализ положительно влиял на бизнес, мы торжествовали. Мы сталкивались с проблемами, работая со сложными деловыми партнерами, и нам помогали наставники, оказывающие поддержку. Хотя этот опыт многому нас научил, истинная ценность заключается в том, чтобы делиться этим опытом с другими.

Цель этой книги — стать руководством по вопросам карьеры в области Data Science. Она описывает путь, который человек пройдет, работая в этом направлении. Мы начнем с азов: расскажем, как получить базовые навыки и понять, что на самом деле представляют собой направления работы в DS. Затем мы объясним, как эту работу получить и освоиться на новом месте. Расскажем, как вырасти в должности и в конечном итоге стать руководителем или уйти в другую компанию. Мы намерены сделать эту книгу ресурсом, к которому дата-сайентисты будут возвращаться на новых этапах своей карьеры.

Поскольку основное внимание в этой книге уделено карьере, мы решили не заострять внимание на технических аспектах Data Science. Мы не будем обсуждать выбор гиперпараметров модели или нюансы пакетов Python. Здесь не будет ни одного уравнения или строчки кода — мы знаем, что об этом уже написано множество замечательных книг. Мы же, напротив, хотели обсудить часто упускаемые из виду, но не менее важные нетехнические знания, которые нужны для достижения успеха.

Мы включили в эту книгу много подробностей из личного опыта уважаемых дата-сайентистов. В конце каждой главы вы найдете интервью с реальными специалистами. Они расскажут, как справлялись с трудностями, рассматриваемыми в главе. Мы были очень рады получить удивительные, подробные и откровенные ответы этих людей и считаем, что их примеры из жизни могут научить гораздо большему, чем любое заявление, которое мы могли бы написать.

При написании этой книги мы намеренно решили сосредоточиться на уроках, которые извлекли, будучи профессионалами в области Data Science, а также общаясь с другими членами сообщества. Иногда мы заявляем о чем-нибудь, с чем не все могут согласиться, например предлагаем всегда писать сопроводительное письмо при поиске работы. Мы решили, что поделиться мнениями, которые, на наш взгляд, будут полезными для дата-сайентистов, важнее, чем пытаться написать что-либо содержащее только объективные истины.

Мы надеемся, что эта книга станет для вас полезным руководством в построении карьеры в области Data Science. Когда мы сами были начинающими специалистами, нам не хватало такой книги. Зато теперь она есть у вас.

Благодарности

Прежде всего хотели бы поблагодарить наших супругов Майкла Берковица (Michael Berkowitz) и Хизер Нолис (Heather Nolis). Без них эта книга не появилась бы (не только потому, что Майкл писал первые черновики некоторых разделов, несмотря на то что он профессиональный игрок в бридж, а вовсе не дата-сайентист, и не потому, что Хизер стремилась заполнить половину книги контентом о машинном обучении).

Хотим поблагодарить сотрудников компании Manning, которые помогли нам пройти этот путь, улучшили книгу и вообще сделали ее выход возможным. Особая благодарность нашему редактору Карен Миллер (Karen Miller), которая помогала нам придерживаться графика и координировала работу.

Спасибо всем редакторам, которые читали рукопись на разных этапах и давали неоценимые подробные отзывы. Вот их имена: Бринджар Смари Бьярнасон (Brynjar Smári Bjarnason), Кристиан Таудал (Christian Thoudahl), Даниэль Берец (Daniel Berecz), Доменико Наппо (Domenico Nappo), Джефф Барто (Geoff Barto), Густаво Гомес (Gustavo Gomes), Хагай Люгер (Hagai Luger), Джеймс Риттер (James Ritter), Джефф Ньюман (Jeff Neumann), Джонатан Твадделл (Jonathan Twaddell), Кшиштоф Енджеевский (Krzysztof Jedrzejewski), Малгожата Родацка (Malgorzata Rodacka), Марио Гизель (Mario Giesel), Нараяна Лалитананд Сурампуди (Narayana Lalitanand Surampudi), Пин Чжао (Ping Zhao), Риккардо Маротти (Riccardo Marotti), Ричард Тобиас (Richard Tobias), Себастьян Пальма Мардонес (Sebastian Palma Mardones), Стив Сассман (Steve Sussman), Тони М. Дубицкий (Tony M. Dubitsky) и Юл Вильямс (Yul Williams). Спасибо также нашим друзьям и членам семьи, которые прочитали книгу и внесли свои предложения: Элин Фарнелл (Elin Farnell), Аманда Листон (Amanda Liston), Кристиан Рой (Christian Roy), Джонатан Гудман (Jonathan Goodman) и Эрик Робинсон (Eric Robinson). Ваш вклад помог оформить эту книгу и сделать ее максимально полезной для наших читателей.

Наконец, хотим поблагодарить всех, кто согласился дать нам интервью: Роберт Чанг (Robert Chang), Рэнди Ау (Randy Au), Джулия Силдж (Julia Silge), Дэвид Робинсон (David Robinson), Джесси Мостипак (Jesse Mostipak), Кристен Керер (Kristen

Kehrer), Райан Уильямс (Ryan Williams), Брук Уотсон Мадубуонву (Brooke Watson Madubuonwu), Джарвис Миллер (Jarvis Miller), Хилари Паркер (Hilary Parker), Хизер Нолис (Heather Nolis), Сейд Сноуден-Акинтунде (Sade Snowden-Akintunde), Мишель Кейм (Michelle Keim), Рене Теате (Renee Teate), Аманда Касари (Amanda Casari) и Анджела Басса (Angela Bassa). Кроме того, мы благодарны тем, кто участвовал в создании примечаний на протяжении всей книги и предлагал вопросы для интервью в приложении: Вики Бойкис (Vicki Boykis), Родриго Фуэнтеальба Картес (Rodrigo Fuentealba Cartes), Густаво Коэльо (Gustavo Coelho), Эмили Барта (Emily Bartha), Трей Кози (Trey Causey), Элин Фарнелл (Elin Farnell), Джефф Аллен (Jeff Allen), Элизабет Хантер (Elizabeth Hunter), Сэм Бэрроуз (Sam Barrows), Решама Шейх (Reshama Shaikh), Габриэлла де Кьерос (Gabriela de Queiroz), Роб Штамм (Rob Stamm), Алекс Хейз (Alex Hayes), Людамила Джанда (Ludamila Janda), Аянти Дж. (Ayanthi G.), Аллан Батлер (Allan Butler), Хизер Нолис (Heather Nolis), Йерун Янссенс (Jeroen Janssens), Эмили Спан (Emily Spahn), Тереза Иофчиу (Tereza Iofciu), Бертил Хатт (Bertil Hatt), Райан Уильямс (Ryan Williams), Питер Болдридж (Peter Baldridge) и Хлинур Хадльгримссон (Hlynur Hallgrímsson). Все эти люди предоставили ценную информацию, и вместе они знают гораздо больше, чем мы.

О книге

Книга «Data Science для карьериста» поможет вам войти в сферу DS и стать профессионалом. В ней рассказывается том, кто такие дата-сайентисты, как получить необходимые навыки и какие шаги нужно предпринять, чтобы устроиться на работу. После трудоустройства эта книга поможет вам понять, как развиваться в своей должности и стать в итоге частью сообщества Data Science, а также дорасти до уровня старшего специалиста. Прочитав ее, вы станете уверенно смотреть на предстоящий карьерный путь.

Для кого эта книга

Эта книга предназначена для людей, которые еще не начали работать в Data Science, но в перспективе рассматривают такую возможность, а также для тех, кто только начал трудиться в этой сфере. Начинающие специалисты получат навыки, которые необходимы, чтобы стать дата-сайентистами, а джуниоры узнают, как повысить свою экспертность. Многие темы в книге вроде прохождения интервью и обсуждения оффера — это полезные ресурсы, к которым стоит возвращаться на любом этапе карьерного пути.

Структура книги

Эта книга разбита на четыре части, посвященные этапам, которые проходит начинающий дата-сайентист. В первой части книги, «Data Science. С чего начать», рассказывается о том, что такое DS и какие навыки нужны для работы в этой сфере:

- В главе 1 вы узнаете о функциях дата-сайентиста, а также о различных должностях с аналогичным названием.
- В главе 2 представлено пять примеров компаний, в которых трудятся дата-сайентисты, и показано, как культура и тип каждой из них влияют на работу.

- Глава 3 описывает различные пути, которые можно выбрать для получения важных для дата-сайентиста навыков.
- Из главы 4 вы узнаете, как создавать проекты и делиться ими для создания портфолио.

Во второй части книги, «Как попасть в Data Science», объясняется весь процесс поиска вакансий:

- В главе 5 рассказывается о поиске вакансий и о том, как понять, ради каких из них стоит стараться.
- В главе 6 мы расскажем, как написать сопроводительное письмо и составить резюме, а затем скорректировать их под каждую конкретную вакансию.
- В главе 7 подробно описывается, как проходит интервью и чего от него следует ожидать.
- Из главы 8 вы узнаете, что делать после того, как получен оффер, и как обсуждать его детали.

В третьей части, «Осваиваемся в Data Science», рассматриваются основные моменты первых месяцев работы:

- В главе 9 рассказывается о том, чего следует ожидать в первые несколько месяцев работы в Data Science, а также о том, как провести это время максимально продуктивно.
- В главе 10 рассматривается процесс проведения анализа, являющегося ключевым компонентом большинства должностей в Data Science.
- Глава 11 фокусируется на внедрении моделей машинного обучения, что является необходимым для специалистов, занимающих инженерные должности.
- В главе 12 объясняется, как общаться со стейкхолдерами, — дата-сайентисты занимаются этим чаще, чем большинство других технических специалистов.

В четвертой части, «Как подняться по карьерной лестнице в Data Science», рассматриваются темы для более опытных специалистов, которые ищут способ профессионально вырасти:

- Из главы 13 вы узнаете, что делать с неудавшимися проектами Data Science.
- В главе 14 показано, как стать частью более широкого сообщества дата-сайентистов с помощью участия в конференциях и разработки открытого исходного кода.
- Глава 15 представляет собой руководство по принятию сложного решения об уходе с должности специалиста Data Science.
- Глава 16 — заключительная; в ней рассказывается о должностях, которые могут получить дата-сайентисты по мере продвижения по карьерной лестнице.

Наконец, в приложении мы собрали для вас более 30 вопросов, которые можно услышать во время интервью, а также предложили примеры хороших ответов. Мы пояснили, какие навыки оцениваются при каждом вопросе и как на них лучше отвечать.

Если вы новичок в области Data Science, то начинайте читать с самого начала, а если вы уже работаете в этой сфере, то переходите сразу к той главе, которая предлагает решение вашей текущей задачи. Несмотря на то что последовательность глав соответствует развитию карьеры в этой сфере, их можно читать в произвольном порядке в соответствии с вашими потребностями.

В конце каждой главы — интервью со специалистами, занятыми в разных индустриях. Они рассказывают, как рассмотренные вопросы коснулись их в работе. Мы выбрали тех специалистов, которые внесли весомый вклад в развитие Data Science и которым пришлось пройти интересный путь прежде, чем стать профессионалами.

От издательства

Карьера в Data Science не зависит от страны, в которой вы живете и учитесь. Чтобы двигаться вперед, необходимо лучше понимать, чего от вас ждет работодатель или хедхантер.

Ваши замечания, предложения, вопросы отправляйте по адресу comp@piter.com (издательство «Питер», компьютерная редакция).

Мы будем рады узнать ваше мнение!

На веб-сайте издательства www.piter.com вы найдете подробную информацию о наших книгах.

Об авторах

Эмили Робинсон

Написала Жаклин Нолис

Эмили Робинсон — блестящий старший дата-сайентист в компании Warby Parker; ранее она работала в DataCamp и Etsy.

Впервые я встретила Эмили на Data Day Texas 2018, когда она была одной из немногих слушательниц моего доклада о Data Science в индустрии. В конце моего выступления она подняла руку и задала прекрасный вопрос. К моему удивлению, через час мы поменялись местами — теперь уже я слушала, как она спокойно проводила восхитительную презентацию, и с нетерпением ждала возможности поднять руку и задать ей вопрос. В тот день я уже поняла, какой она трудолюбивый и умный специалист. Несколько месяцев спустя, когда пришло время искать соавтора для моей книги, Эмили Робинсон была первым кандидатом в списке на

эту роль. Отправляя ей электронное письмо, я думала, что мне, скорее всего, откажут: она, пожалуй, была «не моего уровня».

Работа с Эмили над этой книгой была сплошным удовольствием. Она очень заботится о трудностях младших специалистов по работе с данными, а еще у нее есть способность четко выделять важное. Она всегда качественно выполняет свою работу и каким-то образом умудряется одновременно писать статьи в блогах. Наблюдая за ней на других конференциях и общественных мероприятиях, я видела, как она общалась со многими дата-сайентистами, каждый из которых чувствовал себя с ней комфортно. Она также является экспертом в области A/B-тестирования и экспериментирования, хотя ясно, что для нее это просто временный этап. При желании она могла бы взять любую другую область DS и стать в ней экспертом.

Единственное, что меня расстраивает, так это то, что я пишу эти слова о ней на финальном этапе создания книги, и, как только мы закончим, возможность сотрудничать с Эмили появится уже у кого-то другого.

Жаклин Нолис

Написала Эмили Робинсон

Когда меня спрашивают о том, стоит ли писать книгу, я всегда отвечаю: «Только если у вас будет соавтор». Но это еще не все. Полный ответ должен быть таким: «Только если у вас будет такой же веселый, душевный, щедрый, умный, опытный и заботливый соавтор, как Жаклин». Я не знаю, каково писать книгу с «нормальным» соавтором, потому что Жаклин всегда была просто потрясающей, и мне невероятно повезло поработать с ней над этим проектом.

На фоне такого образованного человека, как Жаклин, вы запросто можете почувствовать себя неловко. У нее есть степень кандидата наук в промышленной инженерии и $100 000 за победу в третьем сезоне телевизионного реалити-шоу «Король ботанов». Жаклин работала директором по аналитике и основала собственное успешное консалтинговое агентство. Она выступает на конференциях по всей стране и регулярно получает приглашения от своей альма-матер приехать и провести карьерные консультации для студентов-математиков (ее специализация). Когда она выступает на онлайн-конференциях, ее забрасывают комплиментами вроде «это лучшее, что я когда-либо слышал», «превосходное выступление», «действительно полезно», «отличная живая презентация». Но Жаклин никогда не дает людям повода чувствовать себя недостойно или плохо из-за того, что они чего-то не знают; наоборот, она любит делать сложные понятия простыми, как, например, в ее презентации «Глубокое обучение — это нетрудно, даю слово».

Ее личная жизнь тоже впечатляет — у нее прекрасный яркий дом в Сиэттле, где она живет со своей подругой, сыном, двумя собаками и тремя кошками. На-

деюсь, однажды она приютит соавтора, чтобы заполнить немного оставшегося места. Она со своей подругой Хизер даже провели презентацию перед аудиторией в тысячу человек об их опыте в использовании R для развертывания моделей машинного обучения в производство T-Mobile. А еще у них, пожалуй, самая милая история знакомства: они встретились на том самом шоу «Король ботанов», где Хизер также была участницей.

Я очень благодарна Жаклин за этот опыт, ведь она могла бы заработать гораздо больше, занимаясь чем-то гораздо менее утомительным, чем написание этой книги вместе со мной. Надеюсь, что наша работа подтолкнет начинающих датасайентистов стать частью сообщества людей, таких же прекрасных, как Жаклин.

Об обложке

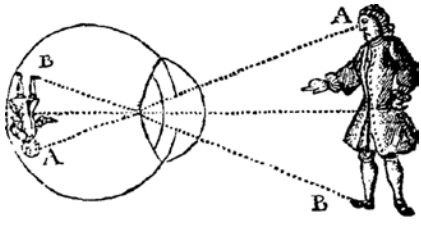
Сен-Совер

Рисунок на обложке книги называется «Femme de l'Aragon», или «Арагонская женщина». Иллюстрация позаимствована из книги Жака Грассе де Сен-Совера (1757–1810) «Костюмы разных стран» (фр. Costumes de Différents Pays), изданной во Франции в 1797 году. Каждая иллюстрация тщательно прорисована и раскрашена вручную. Богатое разнообразие коллекции Сен-Совера ярко отражает то, насколько далекими в культурном плане были города и регионы еще каких-то 200 лет назад. Будучи изолированными, люди говорили на разных языках и диалектах. На улицах городов и деревень по одежде можно было легко определить статус человека, его место жительства и род занятий.

С тех пор манера одеваться сильно изменилась, а разница между регионами, ранее такая заметная, практически исчезла. Сегодня различать жителей разных континентов стало гораздо труднее, не говоря уже о разных городах, регионах или странах. Возможно, мы отказались от культурного многообразия в пользу более разносторонней личной жизни — и уж точно в пользу более разнообразной и быстрой технологической жизни.

В то время когда большинство книг о компьютерах так похожи, издательство Manning отмечает изобретательность и инициативность компьютерного бизнеса с помощью книжных обложек, основанных на богатом разнообразии жизни регионов двухсотлетней давности, оживающей благодаря иллюстрациям Грассе де Сен-Совера.

Часть 1 Data Science. С чего начать

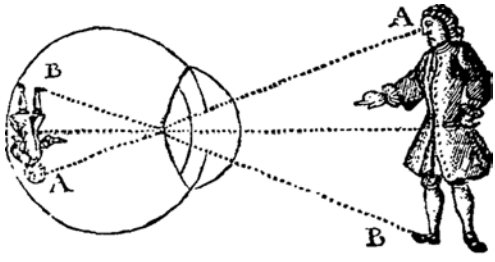


Если вы загуглите «как стать специалистом Data Science», перед вами, скорее всего, появится обширный список, содержащий навыки от статистического моделирования до программирования на Python, а также информация об эффективном общении и проведении презентаций. В одной вакансии может описываться роль, схожая с ролью специалиста по статистике, в то время как другой работодатель ищет кого-то с дипломом магистра информатики. Интернет вам предложит различные варианты приобретения нужных навыков — от возвращения в университет на магистерскую программу до прохождения учебного курса или практики анализа данных на текущем месте работы. В совокупности все эти способы могут показаться непреодолимыми, особенно для тех, кто еще до конца даже не определился с решением стать дата-сайентистом.

Для вас есть хорошая новость: не существует ни одного специалиста по Data Science, который обладал бы всеми этими навыками. У дата-сайентистов есть общий фундамент знаний, но каждый из них специализируется в конкретной области, причем настолько, что многие не смогут поменяться обязанностями. Первая часть этой книги призвана помочь вам разобраться во всех этих специализациях и в том, как принимать наилучшие решения для старта вашей карьеры. К концу у вас будет понимание того, как начать поиск работы.

В главе 1 раскрываются основы работы в Data Science, включая описание необходимых навыков и различных специализаций. В главе 2 подробно рассказывается о роли дата-сайентиста и о пяти типах компаний — это поможет вам лучше понять, на что будет похожа реальная работа. В главе 3 описываются различные пути приобретения навыков, а также преимущества и недостатки каждого из них. Из главы 4 вы узнаете, как создать портфолио как для практического опыта, так и для потенциальных работодателей.

1



Что такое Data Science?

В этой главе

- Три основных направления Data Science.
- Разные типы должностей в области Data Science.

«Самая сексуальная работа XXI века», «Лучшая работа в Америке»... Дата-сайентист — должность, названия которой даже не существовало до 2008 года, теперь является одной из самых востребованных среди соискателей, а работодатели не могут найти достаточное число подобных сотрудников. У такого ажиотажа есть веская причина: Data Science — это быстро развивающаяся область, медианная базовая зарплата специалистов которой в США в 2019 году составила более $100 000 (http://mng.bz/XpMp). В хорошей компании дата-сайентисты пользуются большой автономией и постоянно изучают что-то новое. Они используют свои знания для решения серьезных задач: например, работают с врачами во время испытаний лекарственных препаратов, помогают спортивной команде в подборе новобранцев или изменяют модель ценообразования для бизнеса по производству виджетов. Наконец, в главе 3 мы поговорим о том, что универсального способа стать дата-сайентистом нет. В эту сферу приходят люди с разным образованием, поэтому вы не ограничены своей бакалаврской специальностью.

Однако не вся работа в сфере DS идеальна. И у компаний, и у соискателей бывают нереалистичные ожидания. Например, компании, плохо знакомые с Data Science, могут считать, будто один человек может решить все их задачи с помощью данных. Когда дата-сайентист наконец принят на работу в такую компанию, он

сталкивается с бесконечным списком дел. Ему могут поручить немедленно внедрить систему машинного обучения, при том что никакие работы по подготовке или очистке данных предварительно не проводились. Иногда случается так, что никто не может ему помочь, направить или хотя бы посочувствовать при возникновении проблем. Мы поговорим об этом подробнее в главах 5 и 7, где расскажем, как не оказаться в не подходящих для новичка компаниях, а в главе 9 посоветуем, что делать, если вы попали в неприятную ситуацию.

С другой стороны, соискатели могут подумать, что им никогда не придется скучать. Они могут рассчитывать на то, что стейкхолдеры будут просто следовать их советам, дата-инженеры смогут в мгновение ока исправить любые проблемы с качеством данных, а сами они получат самые быстрые вычислительные ресурсы из возможных для реализации своих моделей. На самом деле дата-сайентисты тратят много времени на очистку и подготовку данных, а также на организацию работы с учетом ожиданий и приоритетов других команд. Проекты не всегда оказываются удачными. Высшее руководство может давать клиентам нереалистичные обещания о работе ваших моделей. Основные обязанности могут заключаться в работе с архаичной системой данных, которую невозможно автоматизировать, — каждую неделю она будет требовать многочасового монотонного труда только на их очистку. Дата-сайентисты могут обнаружить множество статистических или технических ошибок с серьезными последствиями в предыдущих расчетах, но они не будут никого интересовать. При этом специалисты настолько перегружены работой, что им просто некогда что-либо исправлять. Дата-сайентиста могут попросить подготовить отчеты, подтверждающие решение руководства, поэтому он может беспокоиться о том, что его уволят в случае, если он предоставит независимое мнение.

Эта книга поможет вам пройти путь становления в качестве специалиста по Data Science и построить карьеру. Мы хотим, чтобы вы получили все преимущества работы в этой сфере и избежали большинства подводных камней. Возможно, вы работаете в смежной области вроде маркетинговой аналитики и подумываете сменить сферу деятельности. Или, может быть, вы уже работаете дата-сайентистом, но ищете новое место работы и полагаете, что подошли к предыдущему процессу поиска недостаточно хорошо. Возможно, вы хотите продолжить карьеру, выступая на конференциях, участвуя в разработке open source, или же стать независимым консультантом. Мы уверены, что, каким бы ни был ваш нынешний уровень, эта книга окажется вам полезной.

В первых четырех главах мы описали, как можно начать путь в Data Science и создать портфолио: так мы попытались решить парадокс, когда опыт можно получить только при изначальном владении практическими навыками. В части 2 мы покажем, как составить сопроводительное письмо и резюме, с которыми вас

точно пригласят на собеседование, и расскажем, как создать сеть контактов для получения рекомендации. Мы также рассмотрим стратегии переговоров, которые, как показывают исследования, позволят вам получить наилучшие условия оффера.

Как дата-сайентисту вам необходимо будет разрабатывать методы анализа, взаимодействовать со стейкхолдерами и, возможно, даже участвовать в развертывании модели в производство. Часть 3 поможет понять, как устроены все эти процессы и как можно самому настроиться на успех. В части 4 вы найдете стратегии, которые помогут вам собраться с силами в тех неизбежных случаях, когда ваш проект терпит крах. А когда вы будете готовы, мы поможем вам решить, как продолжать свою карьеру — стать менеджером, остаться исполнителем или даже стать независимым консультантом.

Однако прежде, чем начать этот путь, вы должны разобраться в том, кто такие дата-сайентисты и какую работу они выполняют. Data Science — это очень широкое поле деятельности, которое включает в себя много направлений, и чем лучше вы понимаете разницу между ними, тем успешнее вы сможете в них развиваться.

1.1. Что такое Data Science?

Data Science (DS) — это практика использования данных, с помощью которой можно попытаться понять и решить реальные задачи. Эта концепция не нова; люди анализируют объемы и тенденции продаж с тех пор, как изобрели ноль. Однако за последнее десятилетие нам стало доступно экспоненциально большее количество данных, чем прежде. Появление компьютеров помогло генерировать их, и только путем машинных вычислений можно обрабатывать так много информации. С помощью компьютерного кода дата-сайентист может преобразовывать или накапливать данные, проводить статистический анализ или тренировать модели машинного обучения (МО). В результате могут быть созданы отчет, информационная панель или модель МО, которую можно будет запустить в непрерывную работу.

Например, если розничная компания не может определиться с местом для нового магазина, она может пригласить дата-сайентиста для проведения соответствующего анализа. Он соберет статистические данные об адресах доставки онлайн-заказов, чтобы понять, где находится потребительский спрос. Специалист также может совмещать выводы о местонахождении клиентов с информацией о демографической ситуации и доходах в этих местах на основании данных переписи населения. С помощью этих датасетов можно найти оптимальное место для нового магазина и создать презентацию Microsoft PowerPoint, чтобы представить рекомендации вице-президенту компании по коммерческой деятельности.

В другой ситуации та же розничная компания захочет увеличить объем онлайн-заказов с помощью персональных рекомендаций во время шоппинга. Дата-

сайентист может загрузить статистику прежних онлайн-заказов и создать модель машинного обучения, которая будет учитывать набор товаров в корзине покупателя и на его основании прогнозировать, что еще ему можно предложить. После этого он будет работать с командой инженеров компании, чтобы каждый раз, когда клиент совершает покупки, новая модель МО показывала рекомендуемые товары.

При попытке освоить сферу DS многие люди сталкиваются с одной проблемой: слишком уж много нужно изучить. Например, программирование (но какой язык?), статистику (но какие методы наиболее важны на практике, а какие в основном академические?), машинное обучение (но чем оно отличается от статистики или ИИ?) и предметную область в той отрасли, в которой они хотят работать (но что, если вы не знаете, где хотите работать?). Кроме того, им необходимо овладеть бизнес-навыками вроде эффективной презентации результатов всем, начиная с других дата-сайентистов и заканчивая генеральным директором. А от вакансий, в которых требуется степень кандидата наук, многолетний опыт работы в Data Science и знание обширного перечня статистических и программных методов, становится только хуже. Как можно приобрести все эти навыки? С чего лучше начать? Что входит в базу?

Если вы изучали различные области DS, возможно, вы знакомы с популярной диаграммой Венна, составленной Дрю Конвеем. По мнению Конвея (на момент создания диаграммы), Data Science находится на пересечении математики и статистики, знаний предметной области и навыков хакинга (то есть программирования). Это изображение часто берется за основу для определения того, кто такой специалист по работе с данными. На наш взгляд, компоненты науки о данных немного отличаются от того, что предложил Дрю Конвей (рис. 1.1).

Мы изменили исходную диаграмму Венна, составленную Конвеем, на треугольник, потому что дело не в том, есть ли у вас навык или нет, а в том, что вы можете развить его лучше, чем другие специалисты. Действительно, все три навыка являются фундаментальными и вам необходимо владеть каждым в определенной степени, но вам не обязательно быть экспертом во всех. Мы поместили в треугольник разные типы специальностей в сфере Data Science. Они не всегда однозначно соответствуют названиям должностей, а даже если и так, то в разных компаниях их названия могут отличаться. Итак, что означает каждый из этих компонентов?

1.1.1. Математика/статистика

На начальном уровне математика и статистика являются базой в работе с данными. Мы разделяем эту базу на три уровня знания:

- *Существование методов*. Если вы не знаете о какой-либо возможности, вы не можете ее использовать. Если дата-сайентисту нужно сгруппировать похожих

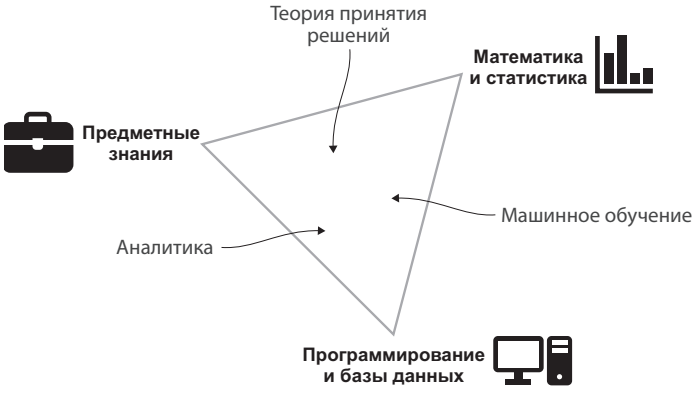


Рис. 1.1. Навыки, которые объединяются в DS, и то, как они сочетаются для выполнения разных функций

клиентов, знание того, что это можно сделать статистическим методом (с помощью *кластерного анализа*), станет первым шагом.

- *Как применять методы*. Специалист по работе с данными должен не просто знать много методов — он должен различать нюансы их применения. Важно писать такой код, где они не только применяются, но и настраиваются. Если дата-сайентист хочет использовать кластеризацию методом *k*-средних, чтобы сгруппировать покупателей, он должен уметь делать это на языке программирования типа R или Python. Также он должен понимать, как настроить параметры метода, например как выбрать количество создаваемых групп.
- *Как выбрать подходящий метод*. В DS используется огромное количество методов, поэтому для дата-сайентиста важно быстро оценить, какой из них будет самым эффективным в каждом случае. В нашем примере с группировкой покупателей, даже если специалист сосредоточился на кластеризации, он может применять десятки различных методов и алгоритмов. Вместо того чтобы перебирать все доступные методы, он должен сразу отбросить бóльшую их часть и сосредоточиться всего на нескольких.

Эти типы навыков постоянно применяются в задачах по работе с данными. Приведем другой пример. Предположим, вы работаете в компании, занимающейся e-commerce. Ваш бизнес-партнер может поинтересоваться, в каких странах у вас самый большой средний чек. Это очень простой вопрос, если у вас есть готовые данные. Но вместо того, чтобы просто предоставить информацию и позволить партнеру делать выводы самостоятельно, вы можете копнуть глубже. Если у вас есть один заказ из страны А на $100 и тысяча заказов из страны Б средней стоимостью

$75, то формально в стране А средний чек выше. Но можете ли вы с уверенностью сказать, что ваш бизнес-партнер должен вложиться в рекламу в стране А, чтобы увеличить количество заказов? Вряд ли. У вас есть только одна единица данных из этой страны, и она может оказаться статистически незначимой. А вот если бы у вас было 500 заказов из страны А, можно было бы протестировать разницу в стоимости заказов. Это значит, что, если бы эти показатели для стран А и Б действительно не различались, вы бы не получили прежний результат. В этом длинном примере дается оценка того, какие подходы были разумными, что следует учитывать и какие результаты были признаны несущественными.

1.1.2. Базы данных и программирование

Программирование и базы данных (БД) основываются на извлечении информации из БД компаний и написании чистого, эффективного, легко настраиваемого кода. Эти навыки во многом схожи с тем, что должен знать разработчик программного обеспечения. Вот только дата-сайентисты должны писать код, который выполняет анализ с неизвестным итогом, а не выдает заранее заданный результат. Стек данных каждой компании уникален, поэтому какой-то определенный набор технических знаний специалисту не нужен. В целом вам нужно уметь получать данные из базы, очищать их, обрабатывать, обобщать, визуализировать и обмениваться ими.

R и Python — основные языки программирования для большинства профессий DS. R берет свое начало в статистике и, как правило, лучше всего подходит для статистического анализа, моделирования, визуализации и составления отчетов. Python создавался как язык для разработки программного обеспечения и в дальнейшем приобрел огромную популярность в обработке данных. Python лучше R справляется с обработкой больших датасетов, проводит машинное обучение и поддерживает алгоритмы, работающие в реальном времени (например, модули рекомендаций в Amazon). Но благодаря вкладу многих участников возможности двух языков сейчас почти равны. Специалисты по работе с данными успешно используют R для создания моделей машинного обучения, запускаемых миллионы раз в неделю, а также делают чистый, презентабельный статистический анализ на Python.

R и Python наиболее популярны для обработки данных по нескольким причинам:

- Они бесплатны, и у них открытый исходный код. Это означает, что он создается многими участниками, а не одной определенной компанией или группой пользователей. В этих языках есть много пакетов, или *библиотек* (готовых блоков кода), которые можно использовать для сбора данных, их обработки, визуализации, статистического анализа и машинного обучения.

- Благодаря большому количеству пользователей каждого из этих языков дата-сайентистам легко найти помощь при возникновении проблем. И хотя в каких-то компаниях до сих пор используют SAS, SPSS, STATA, MATLAB или другие платные приложения, многие из них начинают переходить в своей работе на R или Python.

Хотя бóльшая часть анализа при обработке данных осуществляется на R или Python, часто приходится извлекать информацию из БД, и здесь на сцену выходит язык SQL. SQL – это язык программирования, который используется в большинстве БД для внутренней обработки данных и извлечения их из базы. Представим для примера дата-сайентиста, которому нужно проанализировать сотни миллионов записей о заказах клиентов компании и спрогнозировать, как со временем будет изменяться ежедневное количество заказов. Для начала он, скорее всего, напишет SQL-запрос для получения количества заказов за каждый день, после чего возьмет полученные данные и запустит статистический прогноз на R или Python. По этой причине SQL очень популярен в Data Science, и без знания этого языка вы далеко не продвинетесь.

Можно ли стать дата-сайентистом без программирования?

С данными можно успешно проделывать много вещей, используя только Excel, Tableau или другие BI-инструменты с графическими интерфейсами. Хотя код в них не пишется, часто заявляется, что этот софт так же функционален, как и программирование на R или Python. На самом деле многие дата-сайентисты действительно порой пользуются этими программами. Но могут ли они быть исчерпывающим набором инструментов? Мы говорим «нет». В реальности компаний, где DS-командам не приходится писать код, очень мало. Но даже если вам повезет оказаться в одной из них, у программирования все же есть ряд плюсов.

Первое преимущество программирования — воспроизводимость. Когда вы пишете код, а не пользуетесь программным обеспечением типа point-and-click, можно повторно запускать его при изменении данных хоть каждый день, хоть через полгода. Это преимущество также связано с контролем версий: вместо того чтобы переименовывать файл каждый раз при изменении кода, можно сохранить один файл и видеть всю его историю.

Второе преимущество — гибкость. Например, если в Tableau нет нужного вам типа графа, вы не сможете его создать. Но с помощью программирования можно написать собственный код, чтобы сделать то, о чем создатели и разработчики программных средств никогда даже не думали.

Третье и последнее преимущество языков с открытым исходным кодом, таких как Python и R, — это вклад в сообщество. Тысячи людей создают *пакеты* и публикуют их в открытом доступе на GitHub и/или CRAN (для R) и pip (для Python). Этот код можно скачать и использовать для решения своих задач. Так вы не зависите от числа функций, предлагаемых одной компанией или группой людей.

Другой ключевой навык — использование *контроля версий* для отслеживания изменений кода. Он позволяет организовать хранение файлов, выполнять откат до предыдущих версий и видеть, кто, когда и какие изменения вносил в файл. Этот навык чрезвычайно важен в Data Science и в разработке программного обеспечения. Например, если кто-то случайно изменил файл и испортил ваш код, вы можете восстановить его или посмотреть, что изменилось.

Безусловно, наиболее популярная система для контроля версий — это Git. Он часто используется вместе с GitHub, веб-службой хостинга для Git. Git позволяет сохранять (*фиксировать*) вносимые изменения, а также видеть всю историю проекта и то, как она менялась с каждой фиксацией. Если два человека по отдельности работают над одним и тем же файлом, Git гарантирует, что чья-либо работа не будет случайно удалена или перезаписана. Если вы захотите поделиться своим кодом или запустить что-то в производство, во многих компаниях вам обязательно потребуется Git, особенно если это компания с сильной командой проектировщиков.

1.1.3. Понимание бизнеса

Любая достаточно развитая технология неотличима от магии.

Артур Чарльз Кларк

У компаний, мягко говоря, разное понимание того, как работает Data Science. Часто руководство просто хочет решить определенную задачу и обращается к своим волшебникам DS. Основной навык, необходимый в Data Science, — это умение преобразовать бизнес-ситуацию в вопрос о данных, найти ответ на их основе и предоставить бизнес-решение. Бизнесмен может спросить: «Почему наши клиенты уходят?» Но у Python нет импортируемого пакета «почему уходят клиенты» — вы сами должны понять, как ответить на этот вопрос с помощью данных.

Понимание бизнеса — это та грань, где ваши идеальные представления о Data Science встречаются с условиями реального мира. Недостаточно просто запросить информацию, не зная, как данные хранятся и обновляются в конкретной компании. Если компания предоставляет услуги по подписке, то где хранятся данные? Что произойдет, если кто-то изменит свою подписку? Обновляется ли строка этого пользователя или в таблицу добавляется еще одна? Нужно ли вам исправить какие-либо ошибки или несоответствия в данных? Если вы не знаете всего этого, вы не сможете дать точный ответ на такой простой вопрос, как: «Сколько у нас было подписчиков на 2 марта 2019 года?»

Понимание бизнеса также помогает задавать правильные вопросы. Когда стейкхолдер спрашивает вас, что делать дальше, вероятно, он имеет в виду: «По-

чему у нас нет больше денег?» Для ответа приходится задавать встречные вопросы. Если вы понимаете основной бизнес (а также вовлеченных лиц), то лучше разбираетесь в ситуации. Вы можете спросить в ответ, по какой линейке продуктов нужны рекомендации, или что-то вроде: «Хотели бы вы видеть большее участие определенного сектора нашей аудитории?»

Исчезнет ли Data Science?

В основе вопроса о том, что будет с Data Science через пару десятилетий, лежат две основные проблемы: автоматизация и перенасыщение рынка труда.

Некоторые этапы процесса обработки данных действительно можно автоматизировать. Автоматическое машинное обучение (AutoML) может сравнивать производительность различных моделей и выполнять определенные части подготовки данных (например, масштабирование переменных). Но эти задачи — лишь малая часть большого процесса. Например, данные часто нужно создавать самостоятельно, поскольку идеально чистыми они бывают очень редко. При этом нужно взаимодействовать с другими людьми, например с UX-специалистами или с инженерами, которые будут проводить опрос или регистрировать действия пользователей.

Что касается пузыря на рынке труда, то хорошим сравнением может послужить разработка программного обеспечения в 1980-х годах. По мере того как компьютеры становились дешевле, быстрее и популярнее, возникали опасения, что вскоре эти машины смогут выполнять все и программисты перестанут быть востребованными. Но все произошло ровно наоборот, и теперь в США работает более 1,2 миллиона разработчиков ПО (http://mng.bz/MOPo). Несмотря на исчезновение таких профессий, как веб-мастер, над разработкой, обслуживанием и улучшением веб-сайтов работает больше людей, чем когда-либо.

Мы полагаем, что в Data Science появится больше специализаций, что может привести к исчезновению самого понятия «дата-сайентист». Но многие компании все еще находятся на ранних стадиях изучения того, как использовать науку о данных, и им предстоит еще много работы в этом направлении.

Другая часть понимания бизнеса — это развитие общих бизнес-навыков вроде умения адаптировать презентации и отчеты для разных аудиторий. Иногда вы будете обсуждать лучшую методологию с кандидатами наук по статистике, а иногда вы будете выступать перед вице-президентом, который не занимался математикой уже 20 лет. Вам нужно донести информацию до слушателей, учитывая их особенности.

Наконец, по мере карьерного роста вы научитесь определять, в каких случаях Data Science может помочь бизнесу. Если вы хотели создать систему прогнозирования, а руководство не поддержало эту идею, можно самому стать частью руководства и решить этот вопрос. Старший дата-сайентист будет искать способы

внедрения машинного обучения, так как знает его возможности и ограничения, а также то, какие виды задач выиграют от автоматизации.

1.2. Различные типы вакансий в Data Science

Комбинировать три основных навыка, необходимых в Data Science (и описанных в разделе 1.1), можно на разных по сути должностях. С нашей точки зрения, эти навыки объединяются тремя основными параметрами: аналитикой, машинным обучением и наукой о принятии решений. Каждая из этих областей служит разным целям компании и дает принципиально разные результаты.

При поиске вакансий в сфере Data Science следует меньше обращать внимание на названия должностей — лучше сконцентрируйтесь на описании обязанностей и на вопросах во время собеседования. Посмотрите на опыт работы людей, занимающихся наукой о данных, например какие должности они раньше занимали и на кого учились. Вы можете обнаружить, что должности людей, которые выполняют схожие функции, называются совершенно по-разному, и наоборот, под одним и тем же названием должности «дата-сайентист» может подразумеваться совершенно разная работа. В этой книге мы поговорим о различных типах вакансий, но помните, что названия в разных компаниях могут отличаться.

1.2.1. Аналитики

Аналитик берет данные и передает их нужным людям. После того как компания установит цели на год, их можно поместить на информационную панель, чтобы руководство могло отслеживать прогресс каждую неделю. Можно также встроить функции, которые позволят менеджерам легко разбивать значения по странам или типам продуктов. Эта работа включает в себя много очистки и подготовки данных и, как правило, меньше работы по их интерпретации. Специалист должен уметь находить и устранять проблемы с качеством данных, однако основные решение по ним принимает бизнес-партнер. Таким образом, задача аналитика — взять данные внутри компании, отформатировать, упорядочить и передать их другим специалистам.

Поскольку должность аналитика не связана со статистикой и машинным обучением, некоторые люди и компании считают, что она выходит за рамки Data Science. Однако для большей части работы вроде создания осмысленных визуализаций и принятия решений о конкретных преобразованиях требуются те же навыки, которые нужны и другим специалистам DS. Например, аналитика могут попросить создать автоматизированную информационную панель, которая показывает изменение количества подписчиков и позволяет фильтровать данные только по подписчикам определенных продуктов или в определенных географических

регионах. Он должен будет найти соответствующие данные в компании, выяснить, как их преобразовать (например, изменив их с ежедневных на еженедельные новые подписки), а затем создать содержательный набор информационных панелей с удобным интерфейсом и ежедневным автоматическим обновлением без ошибок.

Короткое правило: аналитик создает *информационные панели и отчеты на основе данных*.

1.2.2. Машинное обучение

Инженер по машинному обучению разрабатывает модели МО и разворачивает их в производство для постоянной работы. Такой специалист может оптимизировать алгоритм ранжирования для результатов поиска на сайте интернет-торговли, создать систему рекомендаций или отслеживать модель в производстве, чтобы убедиться, что ее производительность не снизилась с момента запуска. Инженер по машинному обучению уделяет меньше времени таким вещам, как создание визуализаций для убеждения других людей в чем-то, и больше сосредоточен на программировании для анализа данных.

Существенное различие между этой ролью и другими заключается в том, что результаты работы в первую очередь предназначены для машин. Например, вы можете создавать модели МО, которые превращаются в интерфейсы прикладного программирования (API) для других устройств. Во многих отношениях вы будете ближе к разработчику программного обеспечения, чем к другим специалистам Data Science. Любому дата-сайентисту полезно следовать передовым методам программирования, а вы как инженер по машинному обучению просто обязаны это делать. Ваш код должен быть производительным, протестированным и написанным так, чтобы другие люди могли с ним работать. Поэтому многие инженеры по машинному обучению имеют опыт работы в области информатики.

Инженера по машинному обучению могут попросить создать модель МО, которая может в реальном времени прогнозировать вероятность оформления онлайн-заказа. Он должен будет найти архивные данные в компании, обучить на них модель МО, преобразовать ее в API, а затем развернуть API, чтобы вебсайт мог запускать модель. Если по какой-либо причине эта модель перестанет работать, для решения проблемы пригласят инженера по машинному обучению.

Короткое правило: инженер по машинному обучению создает *модели, которые работают непрерывно*.

1.2.3. Теория принятия решений

Специалист по принятию решений превращает необработанные данные компании в информацию, которая помогает руководству определяться с дальнейшими дей-

ствиями. Для этой работы нужно хорошо владеть различными математическими и статистическими методами и процессами принятия бизнес-решений. Кроме того, специалисты по принятию решений должны уметь создавать убедительные визуализации и таблицы, чтобы люди, не имеющие технических знаний, понимали их анализ. Хотя они много программируют, обычно их код одноразовый — он нужен только для конкретного анализа. Поэтому неэффективный или сложный в поддержке код просто сходит им с рук.

Специалист по принятию решений должен понимать потребности других людей в компании и находить способы выдавать нужную информацию. Например, директор по маркетингу может попросить его помочь определить, какие типы продуктов следует выделить в праздничном каталоге компании. Специалист по принятию решений может исследовать, какие продукты хорошо продавались и без каталога, договориться с командой по user research о проведении опроса и использовать принципы поведенческой психологии, чтобы провести анализ и предложить подходящие варианты. Результатом, скорее всего, будет презентация или отчет PowerPoint, который будет представлен продакт-менеджерам, вице-президентам и другим бизнесменам.

Специалист по принятию решений часто использует знания в области статистики, чтобы помочь компании делать выбор в условиях неопределенности. Например, он может отвечать за управление системой экспериментальной аналитики в компании. Многие компании проводят онлайн-эксперименты или A/B-тестирование, чтобы оценить эффективность изменений. Это изменение может быть простым, например добавление новой кнопки, или сложным, включающим изменение системы ранжирования результатов поиска или полное изменение дизайна страницы. Во время A/B-тестирования посетителям случайным образом предлагается одно из двух или нескольких условий, например *контрольная* группа использует старую версию домашней страницы, а *экспериментальная* — новую версию. По окончании эксперимента действия посетителей из двух групп сравнивают между собой.

Из-за случайности показатели в контрольной и экспериментальной группах редко совпадают. Предположим, вы подбрасываете две монеты и одна выпадает орлом 52 раза из 100, а другая — 49 раз из 100. Можете ли вы сделать вывод, что первая монета имеет склонность выпадать орлом? Конечно, нет! Но бизнес-партнер может посмотреть на эксперимент, увидеть, что коэффициент конверсии составляет 5,4 % в контрольной группе и 5,6 % в экспериментальной, и объявить последнюю успешной. Специалист по принятию решений помогает интерпретировать данные, применять передовые методы разработки экспериментов и так далее.

Короткое правило: специалист по принятию решений создает анализ, на основе которого дает *рекомендации*.

1.2.4. Смежные специальности

Хотя три специализации, о которых мы писали в предыдущих разделах, — это основа работы в Data Science, также бывает несколько других отдельных должностей, которые выходят за рамки этих категорий. Мы перечислим их здесь, потому что разбираться в существующих направлениях полезно и, возможно, вам предстоит сотрудничество с такими специалистами. Тем не менее если вы бы хотели заниматься чем-то из нижеописанного, эта книга может быть для вас менее актуальной.

БИЗНЕС-АНАЛИТИК

Бизнес-аналитик занимается чем-то похожим на работу аналитика, но, как правило, использует меньше статистических знаний и навыков программирования. Его инструментом, вероятнее всего, будет Excel, а не Python, и он может вообще не создавать статистические модели. Хотя его функция аналогична функции аналитика, он выдает менее сложные результаты, поскольку используемые им программные средства и методы ограничены.

Если вы хотите заниматься машинным обучением, программированием или применением статистических методов, должность бизнес-аналитика может вас разочаровать, потому что не даст вам этих навыков. Кроме того, эта работа обычно оплачивается хуже, чем должности в Data Science, и считается менее престижной. Но она может стать хорошим стартом на пути к DS, особенно если у вас нет опыта работы с данными в бизнес-среде. Если вы хотите начать с роли бизнес-аналитика и вырасти до дата-сайентиста, ищите вакансии, где говорится о возможности получить необходимые для вас навыки, например в программировании на R или Python.

ИНЖЕНЕР ДАННЫХ

Инженер данных занимается хранением данных в БД и обеспечением доступа к ним. Он не составляет отчеты, не проводит анализ и не разрабатывает модели; вместо этого он аккуратно хранит и форматирует данные в хорошо структурированных базах для других специалистов. Инженеру данных могут поручить хранение записей о клиентах в крупномасштабной облачной базе и добавление в нее новых таблиц по запросу.

Инженеры данных существенно отличаются от дата-сайентистов — они даже более редкие и востребованные специалисты. Такой сотрудник может помочь создать серверные компоненты данных внутренней экспериментальной системы компании и обновить поток обработки данных, когда задачи начинают занимать слишком много времени. Другие специалисты разрабатывают и отслеживают па-

кетные среды и потоковую передачу, управляя данными на всех этапах от сбора до обработки и хранения.

Если вас интересует инженерия данных, вам потребуются глубокие знания в области информатики; многие инженеры данных — это бывшие инженеры-программисты.

Вики Бойкис (Vicki Boykis): дано ли каждому стать дата-сайентистом?

Учитывая весь оптимизм (и большие потенциальные зарплаты, о которых пишут в новостях) в отношении Data Science, легко понять, почему эта сфера дает привлекательные возможности для карьерного роста, особенно если учесть, что диапазон и количество должностей в DS продолжают расти. Однако начинающему специалисту важно иметь реалистичное и детальное представление о том, как будет развиваться рынок Data Science в ближайшую пару лет, и в соответствии с этим корректировать свои решения.

Сегодня на сферу науки о данных влияет несколько основных тенденций. Во-первых, Data Science как область знаний существует уже десять лет и за это время прошла через ранние стадии цикла хайпа: ажиотаж в СМИ, быстрое внедрение и консолидация. Вокруг DS было много шума, ее обсуждали в медиапространстве, внедряли компании Кремниевой долины и не только, и сейчас мы находимся на этапе быстрого развития области в крупных компаниях и стандартизации таких программных средств обработки данных, как Spark и AutoML.

Во-вторых, в результате быстрого развития отрасли возник избыток новых специалистов, пришедших после изучения новых программ в университетах, буткемпах или на онлайн-курсах. Число кандидатов на любую должность в области Data Science, особенно на начальном уровне, выросло с 20 человек на место до 100 или более. Теперь нередко можно увидеть даже 500 резюме на одну вакансию.

В-третьих, стандартизация наборов программных средств, обеспеченность рабочей силой и спрос на специалистов с опытом работы привели к изменениям в порядке распределения рабочих мест и к созданию иерархии должностей и функциональных обязанностей в Data Science. Например, в одной компании дата-сайентист может заниматься созданием моделей, а в другой — главным образом выполнением анализа SQL, что соответствует, скорее, должности аналитика.

Для тех, кто хочет прийти в Data Science с нуля, это означает несколько вещей. Во-первых, и это самое важное, они увидят, что рынок труда наполнен конкурентами. Особенно это касается тех, кто, в принципе, только начинает работать (например, выпускников колледжей), либо тех, кто пришел в отрасль из какой-либо другой сферы и конкурирует за место с тысячами таких же соискателей. Во-вторых, они могут претендовать на вакансии, которые не совсем соответствуют тому образу Data Science, который создается в СМИ, будто это исключительно написание и внедрение алгоритмов.

Учитывая эти тенденции, важно понимать, что изначально может быть непросто выделиться среди других кандидатов и попасть на финальный этап собеседования. И хотя стратегии, приведенные в этой книге, могут показаться сложными, они помогут вам привлечь внимание, а это необходимо в сложившихся условиях высокой конкуренции.

ИНЖЕНЕР-ИССЛЕДОВАТЕЛЬ

Ученый-исследователь разрабатывает и внедряет новые программные средства, алгоритмы и методологии, которые часто используются другими дата-сайентистами в компании. Такие должности почти всегда требуют наличия кандидатской степени, обычно в области информатики, статистики, количественных социальных наук или в смежных направлениях. Ученому-исследователю может потребоваться несколько недель, чтобы изучить и испытать методы повышения эффективности онлайн-экспериментов, повысить точность распознавания изображений в беспилотных автомобилях на 1 % или создать новый алгоритм глубокого обучения. Он даже может тратить время на написание исследовательских работ, которые будут редко использоваться в компании, но помогут поднять ее престиж и (в идеале) продвинуться в этой области. Поскольку эти должности требуют очень специфического опыта, мы не будем уделять им особого внимания в этой книге.

1.3. Выбор пути

В главе 3 мы рассмотрим несколько способов обучиться работе с данными, опишем преимущества и недостатки каждого из них, а также дадим несколько советов по выбору пути, подходящего именно вам. На этом этапе было бы неплохо задуматься, в каком направлении Data Science вы хотите специализироваться. Какой опыт у вас уже есть? Мы видели дата-сайентистов, которые в прошлом были инженерами, профессорами психологии, менеджерами по маркетингу, студентами программ статистики и социальными работниками. Часто знания, полученные в других профессиях и академических областях, могут помочь вам лучше справляться с работой в DS. Если вы уже работаете с данными, подумайте, в какой части треугольника вы находитесь. Довольны ли вы текущим положением? Хотите ли переключиться на другой тип работы в Data Science? Смена специализации зачастую вполне доступна.

1.4. Интервью с Робертом Чангом, дата-сайентистом из Airbnb

Роберт Чанг (Robert Chang) — дата-сайентист в Airbnb, который работает над продуктом Airbnb Plus. Ранее он занимался аналитикой продуктов, создавал конвейеры данных и модели, проводил эксперименты в «Команде роста» (Growth team) Twitter. Роберт ведет блог об инженерии данных, дает советы новичкам, а также рассказывает о работе в Airbnb и Twitter на странице https://medium.com/@rchang.

Расскажите о вашем первом опыте в Data Science.

Моей первой работой был анализ данных в The Washington Post. Еще в 2012 году я был готов оставить учебу и уйти в эту сферу, но не знал, чем именно хочу заниматься. Я надеялся стать специалистом по визуализации данных, так как был впечатлен работой в The New York Times. Когда я пошел на ярмарку вакансий в вузе и увидел, что в The Washington Post требуются сотрудники, я наивно предположил, что они, скорее всего, делают то же самое, что и The New York Times. Я подал заявку и получил работу, не особо вдаваясь в детали.

Если вам нужен пример того, как не следует начинать карьеру в Data Science, возьмите мой случай! Я получил работу в надежде заниматься либо визуализацией данных, либо моделированием, но очень быстро понял, что, скорее, выполняю обязанности инженера данных. Бóльшая часть моих задач заключалась в создании конвейеров ETL (извлечение, преобразование, загрузка), повторном запуске скриптов SQL и попытках обеспечить запуск отчетов, чтобы можно было представлять ключевые показатели руководству. Тогда я пережил это очень болезненно; я понял, что то, чем мне хотелось заниматься, не соответствовало тому, что было нужно компании, и в конце концов уволился.

Но в последующие годы работы в Twitter и Airbnb я понял, что столкнулся с нормой, а не исключением. При работе с данными их нужно наращивать слой за слоем. Моника Рогати (Monica Rogati) опубликовала знаменитую статью об иерархии потребностей Data Science, попав в самую точку (http://mng.bz/ad0o). Но в то время мне не хватало опыта, чтобы оценить, как в действительности устроена работа в этой сфере.

На что следует обращать внимание при поиске работы в Data Science?

При поиске вакансий вам следует обращать внимание на состоянии инфраструктуры данных в компании. Если вы устроитесь в организацию, где куча сырых данных даже не размещена в хранилище, то уйдут месяцы или даже годы, прежде чем вы займетесь чем-то интересным вроде аналитики, экспериментов или машинного обучения. Если вы на такое не рассчитываете, то этап развития компании совершенно не будет соответствовать тому вкладу, который вы хотите внести в организацию.

Чтобы оценить ситуацию, можно задать вопросы вроде: «Есть ли у вас команда по созданию инфраструктуры данных?», «Как давно она создана?», «На что похож стек данных?», «Есть ли у вас команда дата-инженеров?», «Как они взаимодействуют с дата-сайентистами?», «Есть ли у вас процесс инструментального анализа логов, построения таблиц данных и помещения их в хранилище при создании нового продукта?» Если всего этого нет, вы станете частью команды, создающей все с нуля; приготовьтесь потратить на это немало времени.

Второе, на что нужно обращать внимание, — это люди. Особенно присмотритесь к трем типам сотрудников. Полагаю, вы не хотите быть первым дата-сайентистом в компании. Тогда вам следует искать команду с опытным руководителем. Он знает, как создать и поддерживать хорошую инфраструктуру и процессы, чтобы работа специалистов была эффективной. Также ищите менеджера, который поддерживает постоянное обучение. Наконец, очень важно, особенно для новичков, работать с техническим руководителем проекта или старшим специалистом по данным, у которого много практического опыта. Именно этот человек помогает вам лучше всего справиться с ежедневными задачами.

Какие навыки нужны дата-сайентисту?

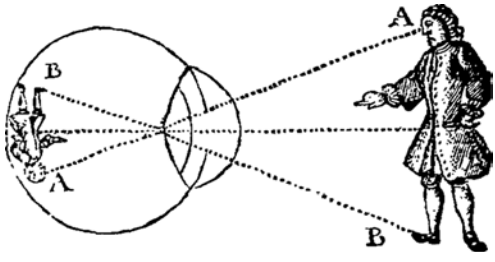
Я думаю, это зависит от того, на какую должность вы претендуете и чего от вас ожидает работодатель. Престижные компании, как правило, задают высокую планку — иногда необоснованно высокую, ведь к ним выстраивается очередь из желающих. Обычно они ищут «единорогов» — тех, кто работает с R или Python, а также отлично разбирается в инженерии данных, проектировании экспериментов, создании конвейеров ETL и моделей с последующим внедрением в производство. Очень уж много требований к кандидатам! Хотя со временем вы можете освоить все эти полезные навыки, не думаю, что они так уж нужны для начала работы в отрасли.

Если вы знаете R или Python и немножко SQL, это уже довольно неплохо для старта. Здорово, если вы можете выучить что-то наперед в целях карьеры, но мне кажется, что это необязательно. Гораздо важнее в принципе любить учиться. У ведущих технологических компаний могут быть более высокие требования, но они нужны скорее не для работы, а для того, чтобы выделить вас среди остальных. Следует различать основные навыки, необходимые для начала карьеры в Data Science, и те, которые неплохо бы иметь сотрудникам топовых компаний.

Итоги

- Набор навыков в Data Science зависит от людей и должностей. Хотя некоторые знания являются фундаментальными, специалисты по работе с данными не обязательно должны быть экспертами во всех смежных областях.
- У работы в Data Science разные направления: предоставление правильных, очищенных данных стейкхолдерам (аналитика), развертывание моделей МО в производство (машинное обучение) и использование данных для принятия решений (теория принятия решений).

2



Типы компаний в Data Science

В этой главе

- Типы компаний, нанимающие дата-сайентистов.
- Плюсы и минусы каждого типа компании.
- Комплекты технологий, которые можно увидеть на разных должностях.

Как уже было сказано в главе 1, в Data Science есть много разных специализаций: инженер-исследователь, инженер по машинному обучению, бизнес-аналитик и другие. Ваши рабочие обязанности будут зависеть от должности, а также от компании, в которую вы устроились. Ее размер, возраст, отрасль — все это влияет на типы проектов, сопутствующие технологии и командную культуру. Умение разбираться в архетипах компаний лучше подготовит вас к поиску работы, будь то ваша первая или очередная должность в Data Science.

Цель этой главы — сформировать у вас представление о повседневной работе некоторых стандартных видов компаний. Мы расскажем о пяти вымышленных фирмах, которым нужны дата-сайентисты. Все эти образы основаны на исследованиях и на нашем собственном опыте. Кроме того, они иллюстрируют основные принципы, которые можно широко применять при поиске работы. Хотя абсолютно одинаковых компаний не существует, знания об этих пяти архетипах поможет лучше понять потенциального работодателя.

Описанные нами стереотипы — не истина в последней инстанции, хоть они и основаны на тенденциях, которые мы наблюдаем в этих отраслях. Есть компании,

которые вообще не соответствуют этим стереотипам, а еще бывает так, что отдельные команды отличаются по своей культуре и организации от остальной фирмы.

Хотя организации в этой главе вымышленные, все остальное написано настоящими дата-сайентистами, работающими в реальных компаниях!

2.1. КИТк: крупная информационно-технологическая компания

- Похожа на: Google, Facebook и Microsoft.
- Возраст компании: 20 лет.
- Количество сотрудников: 80 000.

КИТк — влиятельная технологическая компания, продающая облачные сервисы и специализированное ПО для повышения производительности — текстовые редакторы, серверное оборудование и бесчисленное количество разовых бизнес-решений. Свое огромное состояние компания использует для финансирования необычных проектов в области исследований и разработок (НИОКР), таких как беспилотные скутеры и технологии виртуальной реальности (VR). Об их исследованиях говорят в новостях, а большинство технических сотрудников — это инженеры, которые постепенно совершенствуют уже имеющиеся продукты, добавляют дополнительные функции, улучшают пользовательский интерфейс и запускают новые версии.

2.1.1. Команда: одна из многих в КИТк

В КИТк около тысячи дата-сайентистов. Они собраны в команды, каждая из которых поддерживает свой продукт или подразделение. Кроме того, специалиста могут направить в отдел другого профиля для всесторонней поддержки. Например, у команд проектировщиков VR-шлемов, маркетологов, специалистов по продвижению VR-шлемов и менеджеров цепочек поставок есть свой дата-сайентист.

Если бы вы стали членом одной из этих команд по анализу данных, то быстро бы адаптировались. В крупных организациях новых сотрудников нанимают еже-

дневно, поэтому в компании должны быть стандартные процессы выдачи ноутбука и обеспечения доступа к данным. Также сотрудников обучают работать со специализированным ПО. В команде вам поручат заниматься анализом данных для конкретной области. Это может быть создание отчетов и диаграмм, которые помогут менеджерам обосновать бюджеты проектов. Вам также могут поручить построение моделей МО — они передаются разработчикам для запуска ПО в продакшен.

Скорее всего, в вашей большой команде будет полно опытных специалистов. Поскольку КИТк — компания крупная и успешная, она может привлекать множество профессионалов. Вы будете работать в большой команде, члены которой нередко работают над практически несвязанными задачами, например один сотрудник может выполнять исследовательский анализ на R для директора, а другой — строить модель МО на Python для соседнего отдела. Размер команды — это и благословение, и проклятие в одном флаконе: вы можете обсудить свои идеи со многими экспертами, но большинство из них, скорее всего, не знакомы с вашими конкретными задачами. Кроме того, в команде есть устоявшаяся иерархия. К специалистам на более высоких должностях, как правило, прислушиваются чаще, потому что они опытнее и в своей профессиональной сфере, и в работе с различными отделами КИТк.

Работа вашей команды — это здоровый баланс между поддержанием деятельности компании (например, составление ежемесячных отчетов и ежеквартальное обновление модели МО) и реализацией новых проектов (например, создание новых прогнозов). Руководитель команды должен искать золотую середину между потоком запросов от других команд, которым результаты нужны в ближайшее время, и желанием взяться за что-то инновационное — не востребованное сейчас, но полезное в долгосрочной перспективе. Крупные финансовые возможности КИТк позволяют компании заниматься инновациями и НИОКР гораздо больше, чем другим организациям. Благодаря этому, в свою очередь, команды охотно работают над новыми интересными проектами в Data Science.

2.1.2. Технология: продвинутая, но неупорядоченная

КИТк — крупная организация. При таких масштабах не избежать использования различных типов технологий между подразделениями. Один отдел может хранить данные о заказах и клиентах в базе Microsoft SQL Server, другой — записывать все в Apache Hive. Мало того, неупорядоченными являются не только технологии хранения данных, но и сами данные. Неупорядоченные технологии хранения — еще полбеды, ведь сами данные тоже ведутся по разным принципам. Одно подразделение индексирует записи о клиентах по номеру телефона, другое — по адресу электронной почты.

У большинства организаций такого же масштаба есть собственный арсенал технологий. Поэтому вам как сотруднику КИТк придется освоить способы работы

с данными, характерные именно для этой компании. Изучение специализированного софта здорово поможет на текущей должности, но не в других фирмах.

Вам как специалисту по данным наверняка понадобится несколько видов инструментов. Поскольку КИТк — компания весьма крупная, она хорошо поддерживает распространенные языки, такие как R и Python. Некоторые команды порой работают с платными языками вроде SAS или SPSS, но это бывает реже. Если вы хотите использовать необычный язык, который нравится вам, но мало кем используется (скажем, Haskell), нужно будет получить согласие руководителя.

Комплекс технологий МО сильно различается в зависимости от отдела. Некоторые группы используют микросервисы и контейнеры для эффективного развертывания моделей, тогда как другие работают с устаревшими производственными системами. Разнообразие стека для развертывания ПО затрудняет подключение к API других команд; единой базы знаний или хотя бы понимания того, что происходит, попросту нет.

2.1.3. Плюсы и минусы КИТк

Быть дата-сайентистом в КИТк означает иметь потрясающую работу в потрясающей компании. А поскольку эта компания технологическая, сотрудники знают, кто такой специалист по данным и что полезного он может сделать. Когда все понимают вашу роль одинаково, это значительно облегчает работу. Если в компании много дата-сайентистов, значит, у вас будет широкий круг поддержки, а также возможность плавно влиться в команду и получить доступ к необходимым ресурсам. Оказаться в затруднении один на один — редкость.

В то же время у наличия толпы специалистов по работе с данными есть свои недостатки. Стек технологий сложен, в нем непросто ориентироваться, потому что создавался он разными людьми и разными способами. Может так случиться, что анализ, который вас попросили воссоздать, написал человек, который уже уволился, да еще и на незнакомом вам языке. Вам будет сложнее выделиться среди множества других специалистов. Кроме того, может быть непросто найти интересный проект, потому что над многими из них уже работают другие люди.

Как устоявшаяся компания КИТк дает больше гарантий занятости. Риск увольнений есть всегда, но работа здесь не похожа на работу в стартапе, где финансирование может прекратиться в любой момент. Кроме того, в крупных компаниях руководители больше склонны искать новых сотрудников, чем увольнять старых, потому что увольнение сложно юридически.

У сотрудников КИТк много специализаций — это одновременно и хорошо, и плохо. Дата-инженеры, архитекторы данных, дата-сайентисты, маркетологи и другие выполняют разные задачи, связанные с Data Science, а значит, вокруг вас будет много людей, которым можно передать работу. Например, создавать

собственную базу данных вас вряд ли заставят. С одной стороны, хорошо иметь возможность делегировать задачи, для которых у вас нет опыта, а с другой — так вы не получите новые навыки.

Еще один минус КИТк — бюрократия. В крупной компании введение новых технологий, поездки на конференции и запуск проектов придется согласовывать с начальством. Хуже того, от проекта, над которым вы работали годами, могут отказаться из-за конфликта между двумя руководителями, а ваш проект может «пострадать от шальной пули». Или, что еще хуже, ваш проект может пасть случайной жертвой конфликта двух руководителей — его могут просто закрыть.

КИТк — отличная компания для дата-сайентистов, которые хотят решать сложные задачи с помощью передовых методов. Это касается и специалистов по принятию решений, планирующих заниматься анализом, и инженеров МО, мечтающих создавать и развертывать модели. У крупных компаний есть масса задач и денег, чтобы пробовать новые вещи. Возможно, вы не сможете самостоятельно принимать важные решения, но будете знать, что внесли в них свой вклад.

Работа в КИТк не подойдет специалистам, которые хотят самостоятельно руководить и принимать решения. В большой компании есть установленные методы, протоколы и модели, которым придется следовать.

2.2. HandbagLOVE: устоявшийся ритейлер

HandbagLOVE

- Похожа на: Payless, Bed Bath & Beyond и Best Buy[1].
- Возраст компании: 45 лет.
- Количество сотрудников: 15 000 (10 000 в розничных магазинах, 5000 в офисах).

HandbagLOVE — это розничная сеть с 250 точками по всей территории США, которая занимается продажей кошельков и клатчей. Здесь трудятся оформители магазинов и специалисты по повышению качества обслуживания клиентов. Компания на рынке уже давно, но новые технологии осваивать не спешит: прошло довольно много времени, прежде чем у нее появились первый веб-сайт и приложение.

В последнее время продажи HandbagLOVE упали, поскольку Amazon и другие интернет-магазины потеснили компанию на рынке. Руководство осознало очевид-

[1] Американские сети магазинов одежды и товаров для дома с низкими ценами. — *Примеч. ред.*

ное и решило улучшить ситуацию с помощью технологий, инвестируя в онлайн-приложение и Amazon Alexa, а также пытаясь использовать накопленные данные. Финансовые аналитики HandbagLOVE уже много лет прекрасно рассчитывают совокупную статистику по заказам и клиентам, но лишь недавно компания подумала о том, чтобы нанять дата-сайентистов для лучшего понимания клиентов.

Новая группа специалистов по анализу данных была создана на базе службы финансовых аналитиков, которые ранее составляли отчеты по показателям эффективности компании в Excel. После дополнительного привлечения дата-сайентистов команда начала создавать более сложные продукты: ежемесячные статистические прогнозы роста клиентов в R, интерактивные информационные панели для лучшего понимания продаж, а также сегментацию, объединяющую клиентов в удобные группы для целей маркетинга.

Даже после создания моделей МО для новых отчетов и анализа HandbagLOVE далека от внедрения их в непрерывный рабочий процесс. Все рекомендации по продуктам на ее веб-сайте и в приложении основаны на продуктах МО от сторонних производителей. В команде по анализу данных надеются изменить ситуацию, но никому не известно, когда это все же произойдет.

2.2.1. Команда: небольшая группа, стремящаяся к росту

Команда полагается на специалистов по созданию отчетов, а не по машинному обучению, потому что оно для них в новинку. Никто не владел современными методами статистики и МО, так что сотрудникам приходилось вникать во все самостоятельно. Прекрасно, когда люди могут в одиночку изучать новые интересующие их техники. Обратная сторона медали — неэффективные или даже неправильные методы: в компании нет экспертов, которые могли бы проверить работу.

HandbagLOVE наметила общие пути продвижения специалистов по работе с данными на руководящие должности. К сожалению, они не подходят для сферы Data Science: это глобальные цели, скопированные из других областей вроде разработки ПО, потому что никто на самом деле не понимает, какие показатели использовать. Планируя повышение, вы должны убедить своего руководителя, что готовы перейти на следующий уровень, и, если повезет, он сможет получить одобрение для вашей кандидатуры. С другой стороны, если команда будет расти, вы быстро станете в ней старшим.

Сотрудников группы Data Science знают хорошо, потому что они делают отчеты и модели для других отделов компании (маркетинг, цепочка поставок, обслуживание клиентов). Команда пользуется уважением в компании и дружит с другими подразделениями. У дата-сайентистов HandbagLOVE гораздо больше полномочий, чем в других компаниях, из-за размера команды и ее влияния внутри

организации. Их встречи с руководителями высшего звена на важных переговорах — обычное дело.

2.2.2. Технология: устаревшие методы, которые начинают меняться

В разговорах о технологиях в HandbagLOVE вы часто слышите фразу: «Ну, мы всегда так делали». Данные о заказах и клиентах хранятся в базе данных Oracle, которая напрямую связана с кассовым аппаратом и за 20 лет ни разу не менялась. Система вышла за пределы своих возможностей и претерпела множество изменений. Тем не менее она все еще работает. Другие данные также собираются и хранятся в центральной базе: информация с веб-сайта, центра обслуживания клиентов, рекламных акций и маркетинговых рассылок. Все эти серверы, которые обслуживает ИТ-команда, располагаются локально (*on-prem*), а не в облаке.

Когда все данные хранятся на одном большом сервере, можно свободно подключаться и объединять их как угодно. И хотя иногда запрос занимает вечность или перегружает систему, обходными путями обычно получается найти рабочий способ. Большинство аналитических операций выполняется на ноутбуке. Более мощный компьютер для обучения моделей получить непросто. У компании нет стека технологий для машинного обучения, потому что нет МО как такового.

2.2.3. Плюсы и минусы HandbagLOVE

Как сотрудник HandbagLOVE вы очень влиятельны и можете делать все, что считаете нужным. Можно предложить создать модель пожизненной ценности клиента, построить ее и использовать в компании и при этом не просить разрешения у кучи людей. Такую свободу дает сочетание размера компании и новизны сферы Data Science. И она того стоит: перед вами открываются невероятные возможности для принятия лучших, на ваш взгляд, решений. С другой стороны, вокруг не так много людей, к кому можно обратиться за помощью. Вы сами несете ответственность за то, чтобы все работало, а также за последствия в случае неудачи.

Стек технологий устарел, и вам придется потратить много времени на поиск обходных решений, что, безусловно, не очень практично. Возможно, вы захотите использовать более новый способ хранения данных или запуска моделей, но не получите технической поддержки. Если вы не можете создать какую-либо новую технологию самостоятельно, вам придется обходиться без нее.

Заработная плата будет ниже, чем в более крупных компаниях, особенно в технологических. У HandbagLOVE просто нет денег, чтобы платить за анализ данных. Кроме того, компания в любом случае не ищет лучших из лучших — ей просто нужны люди, которые умеют делать базовые вещи. При этом зарплата не

будет совсем уж низкой: безусловно, она будет намного выше, чем у большинства сотрудников с тем же сроком работы.

HandbagLOVE подходит для дата-сайентистов, которым нравится принимать собственные решения, но при этом не нужны передовые технологии. Если вы не против использовать стандартные статистические методы и составлять рутинные отчеты, HandbagLOVE станет хорошим местом для развития карьеры. Если же вы хотите связаться с новейшими технологиями МО, то таких проектов будет крайне мало; кроме того, в компании практически не будет людей, которые поймут хоть что-то из того, о чем вы говорите.

2.3. Seg-Metra: стартап на ранней стадии

- Похожа на: тысячи неудачных стартапов, о которых вы даже не слышали.
- Возраст компании: 3 года.
- Количество сотрудников: 50.

Seg-Metra — молодая компания, чей продукт помогает клиентам оптимизировать веб-сайты с помощью кастомизации уникальных сегментов плкупателей. В начале своей короткой истории Seg-Metra привлекла нескольких известных клиентов к использованию своих технологий и благодаря этому смогла получить больше финансирования от венчурных капиталистов. Теперь, имея миллионы долларов, компания хочет быстро увеличить размеры и улучшить продукт.

Самое крупное усовершенствование, которое основатели компании предлагали инвесторам, — добавление в продукт базовых методов машинного обучения, что было представлено как «передовой ИИ». Получив новое финансирование, основатели компании ищут инженеров МО для реализации задуманного. Им также нужны специалисты по принятию решений для составления отчетности об использовании продукта, чтобы лучше понять, как его оптимизировать.

2.3.1. Команда (какая еще команда?)

Новый дата-сайентист вполне может оказаться первым в компании. Или же стать одним из первых и подчиняться, скорее всего, тому, кого взяли раньше всех. Поскольку команда новая, протоколов практически не будет — никаких устоявшихся языков программирования, практик, способов хранения кода или официальных совещаний.

Именно тот дата-сайентист, которого взяли первым, будет отдавать все распоряжения. Скорее всего, культура команды будет зависеть от его личностных качеств. Если этот человек открыт для обсуждения и доверяет другим членам команды, то они смогут принимать решения вместе, например обсуждать, какой язык использовать. Если этот человек привык все контролировать и не готов прислушиваться к мнению других, он будет принимать такие решения самостоятельно.

В такой неструктурированной среде может вырасти очень сплоченный коллектив. Команда Data Science всеми силами пытается заставить работать новые технологии, методы и программные средства, и в результате формируются глубокие связи и дружба. С другой стороны, те, у кого нет власти, могут испытывать огромное эмоциональное насилие со стороны руководства, а поскольку компания небольшая, никто не понесет за это ответственности. Независимо от того, как именно будет развиваться компания Seg-Metra, специалистов по работе с данными здесь ждет непростое время.

Работа команды может захватывать или раздражать — каждый день по-разному. Часто дата-сайентисты проводят анализ впервые, например делают первую попытку использовать данные о покупках для сегментации клиентов или развертывают первую нейронную сеть. Аналитические и инженерные задачи, которые решаются впервые, захватывают дух, ведь это неизведанная территория внутри компании, а специалисты по работе с данными становятся первопроходцами. Иногда работа может быть изнурительной, например когда уже пора предоставить инвестору готовую демоверсию, а модель все еще не сходится. Даже если у компании есть данные, сама инфраструктура может быть настолько запутана, что их просто невозможно использовать. Несмотря на хаотичность работы, выполнение всех этих задач в Seg-Metra помогает дата-сайентистам очень быстро освоить множество навыков.

2.3.2. Технология: передовые методы, собранные воедино

Поскольку Seg-Metra — молодая компания, ей не приходится поддерживать устаревшие технологии. Кроме того, хочется произвести впечатление на инвесторов, а сделать это гораздо проще, когда располагаешь эффектным стеком технологий. Поэтому Seg-Metra использует самые современные и лучшие методы разработ-

ки ПО, хранения и сбора данных, а также анализа и отчетности. Информация хранится в современных облачных сервисах: локально ничего не делается. Дата-сайентисты подключаются напрямую к этим базам и создают модели нейронных сетей МО на крупных экземплярах виртуальных машин Amazon Web Services (AWS) с обработкой графическим процессором. Эти модели развертываются с помощью современных методов программной инженерии.

На первый взгляд технологический комплекс, безусловно, впечатляет. Но компания настолько молода и так быстро растет, что у нее постоянно возникают проблемы с совместной работой различных технологий. Когда специалисты вдруг замечают, что в облаке нет данных, им приходится ждать, пока загруженный задачами дата-инженер решит эту проблему (повезло вообще, если он есть). Было бы здорово, если бы у Seg-Metra была специальная команда разработчиков DevOps для поддержки всего в рабочем состоянии, но пока что бюджет распределен иначе. Кроме того, технологию внедрили так быстро, что даже молодой компании сложно контролировать все ее процессы.

2.3.3. Плюсы и минусы Seg-Metra

В растущем стартапе Seg-Metra много привлекательного. Благодаря росту компании появляются всевозможные интересные задачи в области анализа данных и среда, где дата-сайентисты вынуждены быстро учиться. На таких должностях можно приобрести навыки, которые помогут быстро начать карьеру в Data Science: например, научиться работать в сжатые сроки, эффективно общаться со специалистами, не занимающимися данными, понимать, когда проект следует продолжать, а когда стоит от него отказаться. Развитие этих навыков, особенно в начале карьеры, может сделать вас гораздо более интересным сотрудником, чем люди, которые работали только в крупных компаниях.

Еще одно преимущество работы в Seg-Metra — возможность работать с новейшими технологиями, что определенно делает процесс приятнее. Ведь очевидно, что новые технологии лучше старых. А ваше резюме станет от этого более впечатляющим. Компании, стремящиеся использовать новые технологии, захотят, чтобы вы им в этом помогли.

Хотя зарплата здесь не такая конкурентоспособная, как в более крупных, особенно в технологических компаниях, эта работа предоставляет опционы на акции, которые в перспективе могут стать чрезвычайно ценными. Если в итоге компания станет публичной или будет продана, эти опционы могут стоить сотни тысяч долларов или больше. К сожалению, вероятность того, что это произойдет, находится где-то между шансами избрания в городской совет и в Конгресс США. Так что этот вариант подходит только любителям азартных игр.

Родриго Фуэнтеальба Картес (Rodrigo Fuentealba Cartes), ведущий дата-сайентист в небольшой государственной консалтинговой компании

Компания, в которой я работаю, предоставляет аналитические, статистические и мобильные решения для государственных учреждений, вооруженных сил и правоохранительных органов, а также для некоторых частных клиентов. Как ведущий дата-сайентист, я единственный, кто отвечает за проекты в области анализа данных во всей компании. У нас нет дата-инженеров, обработчиков данных или каких-либо других должностей, потому что этот отдел появился относительно недавно. Зато у нас есть администраторы баз данных, разработчики ПО и системные интеграторы, а я совмещаю функции архитектора системы/программного обеспечения и разработчика открытого исходного кода. Это может показаться странным и создает определенные сложности, но я справляюсь на удивление хорошо.

Расскажу вам одну любопытную историю из своего опыта. Я работал в проекте, где использовалась архивная информация о многих параметрах окружающей среды, таких как ежедневные погодные условия. Из-за отсутствия на исследуемой территории установленных метеостанций нам не хватало критически необходимых данных. Проект оказался под угрозой, и заказчик решил закрыть его через неделю, если сотрудники не смогут найти информацию.

Я решил прилететь в этот район и опросить нескольких рыбаков. Я спросил, откуда они узнавали, что выходить под парусом безопасно. Они сказали, что обычно отправляют корабль, который передает погодные условия по радио. Я отправился на радиостанцию и нашел у них записи сообщений с 1974 года. Дальше я внедрил алгоритм, который мог распознавать рукописные заметки и извлекать нужную информацию, а затем реализовал конвейер обработки с использованием естественного языка, который мог анализировать строки. Благодаря моему приезду и обнаружению этих необычных данных проект был спасен.

Густаво Коэльо (Gustavo Coelho), руководитель небольшого стартапа по анализу данных

Последние одиннадцать месяцев я работаю в относительно новом стартапе, который специализируется на применении ИИ в управлении персоналом. Мы прогнозируем будущие результаты кандидатов или вероятность того, что их наймет определенная компания. Цель прогноза заключается в том, чтобы ускорить процесс найма. Мы в значительной степени полагаемся на снижение смещения в моделях. Это небольшая компания: у нас работает одиннадцать сотрудников, а команда по работе с данными состоит из пяти человек, включая меня. Вся компания стремится помочь нам внедрить готовые модели в производство.

Работа в небольшом стартапе позволяет мне ежедневно изучать и применять новые концепции. Мне нравится решать, как лучше всего настроить процессы обработки данных, чтобы мы могли масштабировать их и дать возможность нашим специалистам сосредоточиться на анализе данных. Подбор персонала — не технологическая область знаний, поэтому более половины усилий в проекте уходит на объяснение клиентам используемых решений и помощь им в освоении новой методологии. И когда мы наконец получаем зеленый свет, дальше много времени уходит на координацию с ИТ-отделом клиента и на их интеграцию в наш конвейер данных.

Один из минусов работы в Seg-Metra — большой объем работы. Рабочая неделя длительностью 50–60 часов не редкость, и компания ожидает, что каждый будет делать все, что может. С точки зрения руководства, если все не будут работать вместе, компания не добьется успеха. Вы и правда станете единственным человеком, использующим все дни отпуска? Эта среда может быть чрезвычайно токсичной, наполненной злоупотреблениями и выгоранием сотрудников.

Компания нестабильна, и, чтобы оставаться на плаву, Seg-Metra рассчитывает на поиск новых клиентов и помощь инвесторов, а это означает низкую гарантию занятости. Вполне возможно, что в какой-то момент в компании примут решение уволить сотрудников или объявить о банкротстве. И все это может произойти без предупреждения. Отсутствие гарантий занятости особенно тяжело для семейных людей, именно поэтому основная масса сотрудников состоит из молодежи. Это также может быть недостатком, если вы хотите работать с более разнообразной и опытной командой.

В целом Seg-Metra дает прекрасную возможность работать с интересными технологиями, быстро научиться многому и иметь небольшой шанс заработать кучу денег. Но для этого требуется выполнять огромный объем работы и находиться в потенциально токсичной среде. Так что эта компания лучше всего подходит для специалистов, которые хотят получить опыт, а затем двигаться дальше.

2.4. Videory: успешный технологический стартап на поздней стадии

- Похожа на: Lyft, Twitter и Airbnb.
- Возраст компании: 8 лет.
- Количество сотрудников: 2000.

Videory — это успешный технологический стартап на поздней стадии, который управляет социальной сетью на основе видео. Пользователи могут загружать 20-секундные видеоролики и делиться ими с сообществом. Компания только получила известность, и все от нее в восторге. По масштабу она и близко не может сравниться с КИТк, но отлично преуспевает в качестве социальной сети и с каждым годом увеличивает клиентскую базу. Videory хорошо разбирается в данных, и наверняка уже несколько лет или даже с самого ее основания в ней работают не-

сколько дата-сайентистов или аналитиков. Команда вплотную занимается анализом и составлением отчетов для поддержки бизнеса, а также созданием моделей МО, чтобы помочь людям работать параллельно со специалистами высокого уровня.

2.4.1. Команда: специализированная, но с разнообразием

Videory все еще находится на той стадии, когда можно собрать всех дата-сайентистов в очень большом конференц-зале. Учитывая размер компании, команда по работе с данными может быть организована по централизованной модели, когда каждый сотрудник подчиняется руководителю группы, а все группы работают в одном большом подразделении организации. Команда Data Science помогает другим отделам, но в основном у нее собственные задачи. Некоторые специалисты даже работают над внутренними долгосрочными научно-исследовательскими проектами, которые не приносят мгновенной выгоды.

Как это часто бывает в компаниях такого масштаба, в Videory есть узкоспециализированные подгруппы. Также есть некоторое разделение между специалистами, занимающимися машинным обучением, статистикой или аналитикой. Компания достаточно мала, так что со временем можно будет переключаться между этими группами. Дата-сайентисты часто общаются, например, на тренингах, ежемесячных встречах и в общем чате Slack — такого не встретишь в компаниях вроде КИТк, которые слишком велики для подобного взаимодействия. При этом подгруппы в своей работе часто используют разные инструменты, а сотрудники с ученой степенью занимаются в основном теоретической работой и публикуют научные статьи.

2.4.2. Технология: стараемся не увязнуть в устаревшем коде

В Videory много устаревшего кода и технологий, а еще, возможно, несколько программных средств, разработанных самостоятельно. Компания пытается не отставать от технологических разработок и планирует перейти на новую систему или улучшить существующую. Как и в большинстве организаций, дата-сайентист почти наверняка отправляет запросы в базу данных SQL. В компании также, вероятно, есть программные средства бизнес-аналитики, потому что многие из тех, кто использует информацию, не связаны с Data Science.

За время работы дата-сайентистом в Videory вы обязательно узнаете что-то новое. У всех подобных компаний есть большие данные и системы для их обработки. Одного SQL будет недостаточно; каждый месяц необходимо обрабатывать миллиарды процессов. Однако вы можете попробовать Hadoop или Spark, если нужно извлечь каки-нибудь пользовательские данные, которые не хранятся в базе SQL.

Анализ данных обычно выполняется на R или Python, а в случае трудностей вам на помощь придет множество экспертов. Машинное обучение разворачивается с помощью современных методов разработки ПО, например микросервисов. Поскольку компания известна как успешный стартап, в ней работает много талантливых людей, использующих передовые методы.

2.4.3. Плюсы и минусы Videory

Размер Videory может быть в самый раз для дата-сайентистов; в компании работает достаточно специалистов, которые смогут наставлять и поддерживать новичка, но при этом команда все еще не очень большая, так что вы со всеми познакомитесь. Направление Data Science важно для компании, а значит, ваша работа может получить признание вице-президентов и, возможно, даже высшего руководства (например, генерального или технического директора). В работе вас будут поддерживать дата-инженеры. Конвейеры данных могут иногда работать медленнее или даже сбоить, но вы не несете ответственности за устранение этих неполадок.

В организации с более чем тысячью сотрудников вам придется иметь дело с неизбежными политическими вопросами. Вас могут заставить генерировать числа, которые хотят видеть другие (например, чтобы выслужиться перед начальством и получить бонус). Вы можете столкнуться с нереалистичными ожиданиями в плане скорости разработки чего-либо. Вы можете сделать что-то на самом деле не нужное для бизнеса просто потому, что об этом попросил ваш руководитель. Иногда вы будете чувствовать, что идете в никуда или что зря потратили время. Организация будет сильно меняться, хотя и не так, как на ранней стадии стартапа; то, что приоритетно в одном квартале, может полностью игнорироваться в следующем.

Хотя другие дата-сайентисты в Videory лучше вас будут разбираться в большинстве тем, связанных с анализом данных, вы можете быстро стать экспертом в конкретной области, например в анализе временны́х рядов. Это может быть прекрасно, если вам нравится менторинг и обучение других, особенно если у вас есть время на то, чтобы больше узнавать о сфере с помощью чтения и различных курсов. Но если вы чувствуете, что никто не может проверить вашу работу или подтолкнуть вас к изучению нового, вам может быть непросто. Всегда будет чему поучиться, но полученные знания не обязательно будут относиться к той области, на которой вы хотите сосредоточиться.

В целом у Videory есть хорошее сочетание некоторых преимуществ других компаний. Компания достаточно крупная, чтобы обеспечить вам окружение из специалистов, способных помочь при необходимости, но в то же время она не настолько большая, чтобы в ней царил бюрократический ад или возникали ситуации, когда функции отделов дублируются. У дата-сайентистов есть много шансов научиться новому, но из-за разделения ролей они не могут попробовать все. Эта компания —

отличное место для специалистов, которые ищут беспроигрышный вариант возможностей роста, число которых при этом не стремится к бесконечности.

Эмили Барта (Emily Bartha), первый дата-сайентист в стартапе среднего размера

Я работаю в стартапе среднего размера, у которого есть продукт, ориентированный на страхование. Как первый дата-сайентист, я помогаю определить стратегию использования данных и внедрения машинного обучения в наш продукт. Я вхожу в группу по обработке данных, поэтому очень тесно сотрудничаю с дата-инженерами, а также с нашим продакт-менеджером, работающим с данными.

Рабочий день начинается с утренней встречи команды дата-сайентистов. Мы обсуждаем запланированные задачи, а также блокировщики и зависимости. Я провожу много времени, копаясь в данных: визуализирую, создаю отчеты и исследую их странности или проблемы с качеством. Я также трачу много времени на документацию. Во время программирования я использую GitHub, как и остальные члены команды инженеров; я прошу их проверять мой код (а я проверяю, что написали они). Также значительную часть дня я провожу на совещаниях или работаю над сторонними задачами совместно с членами своей команды.

После работы в крупных организациях сейчас я с удовольствием работаю в небольшой! Здесь дают много свободы для проявления инициативы. Если вы хотите воплотить свою идею, никто не встанет у вас на пути. Ищите компанию, которая уже инвестировала в инженерию данных. Когда меня наняли, в команде уже было несколько дата-инженеров, а также стратегия для инструментариев управления, сбора и хранения данных. В небольшой компании все постоянно меняется, смещаются приоритеты, поэтому важно уметь адаптироваться. Людям, которым нравится глубоко погружаться в проект и работать над ним месяцами, может не понравиться работа в стартапе, потому что там часто требуется разрабатывать приемлемые решения и сразу переходить к следующей задаче.

2.5. Global Aerospace Dynamics: гигантский государственный подрядчик

GLOBAL AEROSPACE DYNAMICS

- Похожа на: Boeing, Raytheon и Lockheed Martin.
- Возраст компании: 50 лет.
- Количество сотрудников: 150 000.

Global Aerospace Dynamics (GAD) — огромная и богатая компания, ежегодно приносящая десятки миллиардов долларов дохода за счет различных государственных контрактов. Компания разрабатывает все: от истребителей и ракет до интеллектуальных светофоров. Филиалы компании разбросаны по стране, большинство из них не взаимодействует друг с другом. GAD существует уже несколько десятилетий, и многие нынешние сотрудники работают там практически с момента основания.

GAD очень неповоротлива, когда дело доходит до Data Science. Большинство инженерных подразделений занимается сбором данных, но им сложно понять, как их использовать в очень регламентированных процессах. Характер работы не допускает наличия багов в коде: он должен быть тщательно протестирован, поэтому идея внедрения модели машинного обучения, которая имеет ограниченную прогнозируемость в реальном времени, в лучшем случае рискованна. В целом темп работы в компании медленный; девиз мира технологий «Двигайся быстро и ломай преграды» — это полная противоположность менталитету GAD.

Учитывая количество статей об искусственном интеллекте, рост сферы машинного обучения и необходимость использования данных для трансформации бизнеса, руководители GAD готовы начать нанимать дата-сайентистов. Они появляются в группах по всей организации и выполняют такие задачи, как анализ инженерных данных для улучшения отчетности, построение моделей МО для внедрения в продукты, и работают в качестве сервисных провайдеров, решая проблемы клиентов GAD.

2.5.1. Команда: дата-сайентист в море инженеров

Хотя конкретные обязанности зависят от того, где именно и над каким проектом GAD трудится специалист, среднестатистический дата-сайентист — это один человек в команде инженеров. В лучшем случае их может быть два или три. Задача этих сотрудников — помогать инженерам с анализом, построением моделей и представлением продукта. Большинство инженеров в команде очень плохо разбираются в Data Science; они помнят регрессии, которые изучали в вузе, но не знают основ сбора данных или конструирования признаков, не разбираются в трудностях валидации модели или в том, как ее развернуть. Мало кто сможет вам помочь, если что-то пойдет не так, но, поскольку в вашей работе мало кто разбирается, вполне вероятно, что никто ничего и не заметит.

Многие из инженеров команды работают здесь более десятка лет, поэтому они хорошо знакомы со спецификой работы в организации. Кроме того, их образ мышления можно свести к фразе: «Мы всегда так делали, зачем что-либо менять?» При подобном подходе реализовать идеи, предложенные дата-сайентистами, крайне трудно. Более медленный характер оборонной индустрии означает, что люди, как

правило, трудятся не так усердно, как в других местах; сотрудники работают по 40 часов в неделю, да и сокращенный день тоже не редкость. В других местах вы бы переживали из-за огромного количества задач, тогда как в GAD причина для стресса — скука и отсутствие работы.

Повышения происходят по одинаковому сценарию, потому что руководители должны соблюдать правила во избежание предвзятого отношения (чтобы на GAD не подали в суд), а еще потому, что так делали на протяжении десятилетий. Повышение во многом зависит от того, сколько лет вы проработали в компании. При чрезвычайном усердии вас могут повысить на ступень на год раньше или выдать чуть большую премию, но вероятность того, что младший дата-сайентист быстро вырастет и станет ведущим, крайне невысока. С другой стороны, компания редко увольняет своих сотрудников.

2.5.2. Технологии: старые, ржавые и с сильными ограничениями системы безопасности

Хотя стек технологий между группами в GAD сильно различается, все они, как правило, относительно старые, локальные (а не облачные) и завалены протоколами безопасности. Поскольку данные касаются характеристик истребителей, для компании важно, чтобы они не попали в чужие руки. За каждую технологию нужно юридически отчитываться на случай, если что-то пойдет не так, поэтому открытый исходный код обычно не приветствуется. Несмотря на то что Microsoft SQL Server стоит дороже, чем, например, PostGRES SQL, GAD с радостью заплатит Microsoft дополнительные деньги, зная, что при обнаружении дыр в системе безопасности можно будет позвонить в компанию и решить проблему.

На практике данные хранятся в базах SQL-сервера, управляемых ИТ-командой, которая вообще настороженно относится к предоставлению разного рода доступа. Дата-сайентистам разрешено использование данных, но они должны запускать Python на специальных серверах с ограниченным доступом к интернету, чтобы ни одна библиотека не утекла за границу. Им также практически невозможно получить разрешение на использование ПО с открытым исходным кодом, что еще больше усложняет работу.

Если код необходимо развернуть в программном комплексе, это обычно делают традиционными способами. GAD только начинает использовать современные методы внедрения МО в производство.

2.5.3. Плюсы и минусы GAD

Работа отдела Data Science медленная, комфортная и надежная — это преимущество GAD. Менее напряженный темп работы означает, что в конце дня вы,

скорее всего, не будете чувствовать себя как выжатый лимон. В процессе у вас часто будет появляться свободное время для чтения профессиональных блогов и статей, и никто не будет ворчать по этому поводу. Мало кто будет дергать вас с вопросами, ведь почти никто в компании не разбирается в основах Data Science. А поскольку GAD — это крупная организация, для которой важна юридическая ответственность, вам придется сильно постараться, чтобы вас уволили.

Недостаток работы в GAD — небольшой по сравнению с другими компаниями шанс получить новые навыки. Скорее всего, вы будете долгие годы заниматься одним и тем же проектом, поэтому необходимые для него технологии и инструменты вскоре станут примитивными. Хуже того, ваши новые навыки будут связаны с устаревшими технологиями, которые не применяются в других учреждениях. А получить повышение в этой компании так же непросто, как и быть уволенным из нее.

GAD — отличное место, если вы найдете команду, которая занимается интересными проектами, и если вы не хотите жить на работе. Многие работают в компании десятилетиями, потому что это удобно и они этим довольны. Но если для движения вперед вам нужны трудности, GAD может не подойти.

Нейтан Мур (Nathan Moore), менеджер по аналитике данных в энергоснабжающей компании

Компания, в которой я работаю, обеспечивает и продает электроэнергию сотням тысяч людей и частично принадлежит государству. В самой компании около тысячи сотрудников, занимающих самые разные должности. В мои обязанности входит исследование и создание прототипов новых источников данных, а также работа со специалистами БД, очистка и документирование текущих источников данных. У нас полно устаревших систем и новых инициатив, так что нам всегда есть чем заняться.

Сейчас мой рабочий день состоит из совещаний, рассмотрений спецификаций для ETL, тестирования новой техники MO, которую я нашел в Twitter, предоставления обратной связи об отчетности, обучения использованию JIRA и Confluence и ответов на множество электронных писем. Раньше я занимался разработкой и оценкой моделей, анализировал данные, если какой-либо процесс давал сбой, и представлял отчеты по отрасли сектора в целом.

Наша компания достаточно крупная, чтобы иметь хорошую команду аналитиков для работы над множеством задач, от ежедневных отчетов до крупных проектов по сегментации клиентов. У меня было много возможностей поработать в разных сферах этого бизнеса, и я провел здесь 11 лет. Поскольку компания владеет активами стоимостью в миллиарды долларов, она не готова рисковать, а изменения здесь происходят медленно. У нас достаточно крупный ИТ-отдел, который может поддерживать повседневные процессы, но любой значительный проект, например обновление системы, означает, что ресурсы для неприоритетных оптимизаций ограничены. Все должно быть обосновано, на работы должен быть выделен бюджет, и к тому же есть определенная политика компании, которую необходимо соблюдать.

2.6. Делаем выводы

Если вы посмотрите на компании, которые размещают вакансии, вы обнаружите, что многие из них похожи на перечисленные в этой главе. По мере поиска работы и прохождения собеседований постарайтесь понять плюсы и минусы сотрудничества с каждой из них (табл. 2.1) — вам это может пригодиться.

Таблица 2.1. Краткий обзор компаний, которые нанимают дата-сайентистов

Критерий	КИТк	HandbagLOVE	Seg-Metra	Videory	GAD
	крупная технологическая компания	ритейлер	стартап	средняя технологическая компания	защита
Бюрократия	Много	Мало	Нет	Немного	Много
Технологии	Сложные	Старые	Неустойчивые	Смешанные	Допотопные
Свобода	Мало	Много	ОЧЕНЬ МНОГО!	Много	Нет
Зарплата	Восхитительная	Приличная	Плохая	Отличная	Приличная
Гарантия занятости	Отличная	Приличная	Плохая	Приличная	Отличная
Перспективы обучения	Много	Есть	Много	Много	Немного

2.7. Интервью с Рэнди Ау, специалистом в области количественного UX Research в Google

Рэнди Ау (Randy Au) работает в команде Google Cloud. Уже более десяти лет он занимается Data Science в области человеческого поведения. В своем блоге https://medium.com/@randy_au он рассказывает, как правильно рассматривать вакансии в стартапах и в других различных типах компаний.

Есть ли принципиальная разница между большими и маленькими компаниями?

Да. Обычно она касается внутренней организации и структуры. Бывает так, что культура компании меняется из-за ее масштаба. В стартапе из 10 человек «все делают все», потому что каждый выполняет разные роли. Когда в компании набирается около 20 сотрудников, появляются специализации: для работы над

конкретными задачами формируются команды из трех-четырех человек. Они могут лучше сконцентрироваться на определенных вещах, и знать компанию досконально уже не обязательно. Если в команде набирается 80–100 человек, она перестает расти. Многие процессы вращаются вокруг нее. Теперь не получится знать всех сотрудников компании. Неясно, кто чем занят, поэтому нужно выстраивать более сложную иерархию для взаимопонимания. Если в команде набирается примерно 150–200 человек, то приходится мириться с бюрократией, иначе разобраться с тем, что происходит в компании, попросту невозможно. А затем есть Google с его 100 000 человек. Там вообще не понятно, чем занимается бóльшая часть сотрудников.

Чем меньше компания, тем больше вероятность взаимодействия со всеми. В компании из 40 человек генеральный директор сидел бы за моим столом, пока мы бы с ним изучали набор данных. В Google такого никогда не случится. Но готовы ли вы к ситуации, типичной для многих стартапов, когда вы создаете автомобиль для «Формулы 1» и одновременно управляете им, а все при этом спорят, нужен ли там вообще руль? Когда вы отвечаете за данные в небольшой компании, не важно, какими методами вы пользуетесь, — вы просто пытаетесь все сжать и извлечь информацию. Нет ничего страшного в отсутствии жестких рамок, когда нужны быстрые решения.

Отличаются ли компании в зависимости от отрасли?

Исторически сложилось так, что математики и дата-сайентисты изначально работали в определенных отраслях. Например, в страховой компании работают актуарии. Эти специалисты были там уже сотню лет и действительно знают свое дело. Если страховые компании привлекают дата-сайентистов, они руководствуются немного другими целями. У них уже есть служба чрезвычайно талантливых статистиков. Такие компании собираются заполнить пробел в больших данных либо же оптимизировать свой веб-сайт или что-то вроде того.

У финансистов также есть давняя традиция привлекать спецов по количественному анализу. Помню, как я однажды провалил собеседование по финансовому анализу, потому что они проверили мой код. Но мне как дата-сайентисту нужно просто убедиться, что мой код работает и выдает правильный ответ; я не слишком задумываюсь о производительности, пока все работает как нужно. А их тест буквально проверил производительность и снял за это баллы. Я подумал: «А, ну да, вы же, ребята, занимаетесь финансами. Понимаю».

Думаю, если вы поговорите со всеми, кто делает что-то в Data Science, то увидите, что подавляющее, но молчаливое большинство — это люди, которые выполняют совершенно непривлекательную, монотонную работу. Я получил невероятное количество откликов на статью об этой сфере в стартапах, где люди говорили: «Да, это моя жизнь». Об этом не упоминают в обсуждениях науки о данных. Это

не крутые вещи вроде «Вот новый блестящий алгоритм, который я применил из arXiv». Не думаю, что применял что-либо из arXiv за двенадцать лет работы. Я все еще использую регрессионный анализ, потому что это действительно работает! Я думаю, что реальность выглядит именно так.

Вы будете чистить данные; не думаю, что даже в фейсбуках и гуглах есть хоть кто-то, кто не чистит данные. Делать это проще, если есть налаженная структура. Но нет, вам все равно придется этим заниматься. Такова правда жизни.

Что вы посоветуете начинающим дата-сайентистам?

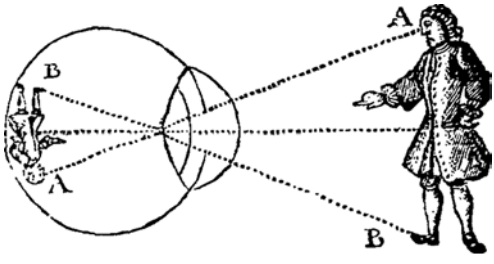
Разбирайтесь в своих данных. Это действительно занимает много времени — от шести месяцев до года или даже больше, если система сложная, но качество данных — это ваша основа основ. Если вы не понимаете данные, с которыми работаете, то однажды сделаете совершенно неправильные выводы. Кто-то скажет: «О, у меня есть подборка уникальных файлов cookie пользователей, посетивших веб-сайт, и их количество равно количеству уникальных посетителей». Но это не так. А как же те, кто использует несколько устройств или браузеров?

Чтобы действительно разбираться в своих данных, нужно подружиться с профессионалами предметной области. Когда я составлял финансовые отчеты, то общался с финансистами в надежде разобраться, что и как называется и что и в каком порядке вычитается по правилам бухгалтерского учета. У вас может быть 50 миллионов посещений с одного IP-адреса, но в сеть с данного компьютера выходят разные пользователи. Вы этого не поймете, а кто-то другой, возможно, поймет.

Итоги

- Многие типы компаний нанимают дата-сайентистов.
- Работа в Data Science различается в основном в зависимости от отрасли, размера, истории и культуры каждой компании.
- Важно понимать, в какой компании вы хотите работать.

3



Приобретение навыков

В этой главе

- Различные способы изучения Data Science.
- Что делает академический курс или буткемп хорошим.
- Как выбрать наиболее подходящий путь.

Теперь, когда вы решили стать дата-сайентистом, нужно приобрести соответствующие навыки! Не переживайте: размышления на тему того, как это лучше сделать, — естественная часть становления всех специалистов. Есть много способов выучиться, начиная с просмотра роликов на YouTube и заканчивая получением диплома, при этом многие будут настаивать, что именно они пошли единственно правильным путем. Хуже того, можно легко впасть в отчаяние от количества материала для изучения: все эти алгоритмы, языки программирования и статистические методы — еще и различные области бизнеса сверху накиньте. Сама мысль обо всем этом приводит в ужас.

Но не все так плохо. Основных методов получения необходимых навыков всего четыре, и у каждого из них есть свои преимущества и недостатки: обычно, если их расписать, можно понять, что вам ближе. К концу этой главы вы сможете научиться разбираться в методах и затем, немного поразмыслив, выбрать наиболее подходящий для себя. Вы справитесь!

Четыре метода получения навыков работы с данными, описанные в этой главе:

- Получение диплома по Data Science или в смежной области.
- Прохождение буткемпа по Data Science (ускоренный курс от 8 до 15 недель).

- Изучение Data Science на текущем месте работы.
- Обучение с помощью онлайн-курсов и книг по Data Science.

Ниже мы рассмотрим все эти методы.

Что делать, если у вас нет диплома?

В этой главе в основном предполагается, что вы окончили вуз, и скорее всего по технической специальности. Если это не ваш случай, не волнуйтесь: бóльшая часть информации по-прежнему актуальна, но читать ее нужно будет с поправкой на этот факт.

Если вы еще не окончили вуз, лучше все же получить степень бакалавра, прежде чем следовать рекомендациям этой главы. Лучший вариант — соответствующее техническое направление, где можно получить некоторые навыки для Data Science, например математика, статистика или информатика. При выборе таких специальностей постарайтесь спланировать программу обучения так, чтобы максимально заполнить пробелы в знаниях. Сейчас некоторые вузы предлагают степень бакалавра в Data Science, что должно сделать из вас весьма подходящую кандидатуру для работодателя. С таким дипломом можно найти работу в этой области сразу после выпуска (особенно если будете следовать указаниям в частях 1 и 2 этой книги). Вы также можете взять на вооружение дополнительные советы из этой главы, например обучаться самостоятельно или заниматься Data Science на первой работе.

Если у вас нетехническая специальность, рекомендации из этой главы остаются в силе. Но закончить магистратуру в Data Science — все же отличная идея, ведь чем дольше длится обучение, тем больше времени у вас будет на подтягивание технических навыков. Возможно, вы захотите получить техническую специальность как второе высшее, но этого следует избегать любой ценой. Получение второго образования чрезвычайно затратно по времени и деньгам; к тому же эти знания можно получить другими способами.

3.1. Получение образования в Data Science

Многие вузы предлагают магистратуру в Data Science по программам со смесью тем из информатики, статистики и бизнеса. Такие программы обычно рассчитаны на два года и стоят от $70 000.[1] Как и в случае с другими направлениями, можно растянуть этот срок и совмещать учебу с работой и/или заниматься онлайн. Хотя многие учебные заведения предлагают образование непосредственно в Data Science, вместо этого вы можете выбрать другое направление — информатику, статистику, бизнес-аналитику, исследования операций или что-то очень близкое к науке о данных.

Преимущество направления Data Science — его всесторонность. Благодаря длительности программы и затраченному времени вы получите все знания,

[1] Здесь речь идет о стандартных программах магистратуры в вузах США. — *Примеч. ред.*

необходимые для начала работы в качестве младшего дата-сайентиста. Курсовые работы и проекты дают опыт использования статистики, методов МО и практического программирования. Даже если в начале обучения вы не будете особо знать программирование, то сможете все наверстать в процессе (хотя возможно, что для этого вам придется пройти пару дополнительных курсов).

Однако у полноценных программ по Data Science есть и несколько недостатков:

- Они очень дорого обходятся, учитывая и стоимость обучения, и шанс не получить ни прибыли, ни опыта работы, пока вы учитесь на дневной форме. Полноценные образовательные программы требуют на порядок больше времени и денег, чем другие варианты. Потратить годы на учебу, чтобы сменить карьеру, — это огромная часть жизни, и если на полпути вы решите, что Data Science не для вас, никто уже не вернет вам эти ресурсы.
- Если вы пришли из смежной сферы — скажем, из разработки ПО — или закончили солидный ряд схожих курсов по программе бакалавриата, то магистратура не даст вам практически ничего нового. То есть из этой длинной программы вы сможете почерпнуть лишь крупицу полезной информации — огромный недостаток, который может раздражать.
- Эти программы создают и читают профессора и преподаватели, в большинстве своем получившие весь свой опыт исключительно в научном сообществе. Материалы, которые они дают, часто существенно отличаются от того, что действительно используется в отрасли. Например, особенно оторванный от практики профессор может пользоваться устаревшими языками программирования вроде SPSS или не разбираться в современных инструментах, таких как управление версиями. Ситуация особенно типична для программ, не относящихся к Data Science. Некоторые университеты, наоборот, привлекают профессионалов отрасли, однако такие люди могут не особо разбираться в преподавании. Без поступления трудно понять, насколько современные методы применяются в обучении. По возможности во время подачи заявления старайтесь поговорить с нынешними или бывшими студентами, чтобы понять, на что похожа программа и поможет ли она в будущей карьере.

3.1.1. Выбор учебного заведения

В огромном количестве образовательных программ по Data Science и так сложно ориентироваться. Еще хуже, если ваш почтовый ящик переполнен флаерами с рекламой различных курсов, а телефон разрывается от назойливых звонков сотрудников приемной комиссии. Лучше всего подавать заявления на 3–10 программ. Подадите слишком мало — рискуете никуда не попасть, слишком много — потратите кучу времени (и денег).

При выборе учебного заведения руководствуйтесь следующими параметрами:

- *Будете ли вы довольны местоположением и образом жизни [очень важно].* Скорее всего, вы будете искать программы магистратуры по всей стране, но студенческая жизнь в Лос-Анджелесе отличается от жизни в Нью-Йорке. Если климат, близость к друзьям или стоимость жизни вам не подходят, то не важно, насколько хороша программа, потому что вы все равно будете недовольны.
- *Какие темы охватывает курс программы* [*важно*]. Поскольку Data Science — новая сфера, университеты могут иметь совершенно разные представления о том, как ее преподавать. Ситуация особенно осложняется тем, на какой кафедре реализуется программа. Если это кафедра информатики, то упор будет делаться на методы и алгоритмы, а если это бизнес-школа, то главными будут прикладные методы и практические кейсы. Проверьте, закрывает ли материал курса пробелы в вашем наборе навыков (см. главу 1).
- *Объем проектной работы в программе* [*важно*]. Чем больше в программе проектов, тем больше вы узнаете о том, как Data Science работает на практике, и тем лучше будете готовы к работе в отрасли. (Проекты подробно рассматриваются в главе 4.) Важные проекты также отлично подходят для включения в резюме, которое поможет пройти стажировку во время обучения в аспирантуре или устроиться на первую работу.
- *Судьба выпускников учебного заведения* [*важно*]. Часто вузы собирают данные о том, где работают студенты после окончания учебы, например какой процент попадает в научные круги или в компании из списка Fortune 500. Эта статистика может быть информативной, но учебные заведения делятся только той, которая представляет их в выгодном свете, даже если показатели обманчивы (ирония в том, что понимание обманчивости показателей — один из навыков, которому вы учитесь на программе Data Science). Чтобы получить объективное представление, попытайтесь по возможности связаться с некоторыми выпускниками программы через LinkedIn. Если хотите работать в крупной корпорации, узнайте, какие компании нанимают выпускников именно этого заведения. Конечно, вы в любом случае можете участвовать в конкурсном отборе при приеме на работу, но, вероятно, вашему отклику будет уделено меньше внимания.
- *Финансирование* [*бывает редко, но очень важно*]. В редких случаях учебные заведения предлагают финансирование для студентов магистратуры, оплачивая их обучение и иногда добавляя стипендию помощникам преподавателей. Если вам предложили стипендию, мы настоятельно рекомендуем согласиться. Не платить за учебу и при этом получать зарплату гораздо приятнее, чем самостоятельно оплачивать счета. Если финансирование предполагает педагогическую прак-

тику, вам придется научиться общаться с большим количеством людей, что пригодится в дальнейшем. Минус такой ситуации в том, что преподавание занимает много времени, а это отвлекает от учебы.

- *Насколько тесно программа взаимодействует с предприятиями в этом регионе* [*довольно важно*]. Если учебное заведение сотрудничает с местными компаниями, особенно с технологическими, значит, оно связано с профессиональным сообществом. В этом случае будет проще получить стажировку или работу, а материалы во время учебы будут интереснее. Кроме того, преподаватели, скорее всего, будут знакомы с практическими методами.
- *Требования к зачислению* [*не очень важно*]. Некоторые учебные заведения требуют, чтобы у поступающих был определенный багаж из пройденных курсов. Для большинства программ это математические дисциплины вроде линейной алгебры и программирование, например введение в Java. Если вам не хватает всего пары курсов, возможно, у вас получится проскочить эти требования или наверстать пробел уже во время учебы на программе. Если же вы не прошли ни один из них или не закончили требуемое направление бакалавриата (например, информатику), возможно, программа вам не подойдет.
- *Престиж учебного заведения* [*совершенно не важно*]. Если речь не идет о невероятно престижных заведениях вроде Стэнфорда или MIT, работодателям все равно, что вы окончили. Престиж имеет значение, в основном если вы планируете работать в научных кругах, а не в промышленности, но тогда вам придется идти в аспирантуру, а не в магистратуру (а также читать другую книгу). Он важен только для формирования крепкого научного сообщества топовых университетов.
- *Ваш научный руководитель* [*очень важно, но...*]. Если программа, которую вы рассматриваете, предполагает написание диплома или диссертации, то у вас будет научный руководитель. Когда его стиль работы и сфера интересов совпадают с вашими, а вдобавок он еще и приятный человек, ваши шансы успешно закончить программу возрастают в разы. К сожалению, до поступления очень сложно определить, кто будет вашим научным руководителем, не говоря уже о его личностных качествах. Так что, несмотря на огромную важность этого критерия, вы вряд ли сможете принимать решение на его основе. А вот если программа полностью основана на выполнении курсовых работ или включает только один финальный проект, личность руководителя не особо важна.

Размышляя над перечнем вузов, попробуйте составить таблицу, в которой будут расписаны их качества по этим критериям. Однако даже со всеми данными на руках сложно объективно ранжировать учебные заведения. Разве можно наверняка решить, что лучше: вуз в ужасном городе, но с хорошими связями в компаниях или вуз в отличном месте, но без проектной работы? Мы советуем

отказаться от идеи найти «лучшее». Вместо этого сгруппируйте варианты по принципам «люблю», «нравится» и «сойдет» и подавайте заявления только в заведения из первых двух групп.

Дистанционные программы магистратуры

Все чаще стали появляться дистанционные программы, где диплом магистра можно получить онлайн и не ходить на лекции непосредственно в университет. Очевидное преимущество этого варианта в том, что проходить онлайн-курсы гораздо удобнее, чем тратить время на дорогу. Кроме того, к онлайн-программам перестали относиться предвзято, как это было на заре их существования, поэтому вам в принципе не следует беспокоиться о ее признании. Недостаток же такого подхода в том, что удаленно гораздо труднее оставаться вовлеченным в программу и материалы. Задавать вопросы профессорам сложнее, зато проще работать в полсилы и отлынивать от выполнения домашних заданий. В некотором смысле удобство онлайн-программы также может быть ее недостатком: у вас остается меньше стимулов для работы. Если вы уверены в своей мотивации и способности удерживать внимание, онлайн-обучение может стать отличным выбором — просто имейте в виду его риски.

3.1.2. Поступление

Для поступления нужно подать документы. Процедуры для магистратуры и бакалавриата похожи, поэтому никаких сложностей возникнуть не должно. Первый шаг — это подача заявления. Осенью университеты обычно публикуют все требования и устанавливают процедуру подачи документов, включая сроки. Для поступления в магистратуру обычно требуется следующее:

- *Мотивационное письмо на 1–2 страницы*, в котором вы должны аргументировать, почему подходите для этой программы. Максимально подробно опишите, каким образом вы могли бы внести хороший вклад в эту программу. Ваши навыки, опыт или примеры работ будут плюсом. Избегайте клише вроде «Я с детства интересуюсь Data Science». Есть множество материалов по написанию хороших эссе, а в выбранном университете даже может быть отдел, помогающий справиться с этой задачей.
- *Выписка об академической успеваемости* из программы бакалавриата, которая подтверждает, что вы соответствуете всем исходным требованиям. На вебсайте университета должны быть инструкции по получению этого документа, но учтите, что обычно за него нужно заплатить, а доставка занимает неделю или больше. Не откладывайте это на последний день!
- *Результаты экзамена Graduate Record Examination* (*GRE*) с проходными баллами по языку и математике. Теоретически GRE по математике должен пока-

заться простым для любого, кто идет в Data Science, ведь математика — основа этой дисциплины. Однако многие не видели сложных математических задач со школы, поэтому лучше подготовиться. Языковая часть может быть сложнее и потребовать серьезной подготовки. Обычно экзамен сдается в конкретном месте, куда вам нужно добраться (а это может быть непросто), поэтому постарайтесь сдать его заранее. Если английский не ваш родной язык, скорее всего, потребуется набрать минимальный балл на экзаменах TOEFL (Test of English as a Foreign Language — тест на знание английского языка как иностранного) или IELTS (International English Language Testing System — международная система оценки знания английского языка).

- *Три рекомендательных письма*, объясняющих, почему вы подходите для этой программы. Это могут быть письма от ваших преподавателей или от начальника, если ваша работа связана с Data Science. В идеале авторы должны рассказать о том, почему вы будете хорошим дата-сайентистом, то есть они должны видеть ваши результаты. Не обращайтесь к преподавателям, которые не могут написать ничего, кроме: «Он получил пятерку на моем семинаре», а также к работодателям, которые мало что могут сказать о вашей работе в технической среде. Если вы студент бакалавриата, который читает эту книгу, возможно, сейчас самое подходящее время, чтобы получше узнать своих преподавателей на консультациях, семинарах и в научных клубах.

Все это требует времени, так что если вы подаете заявления сразу в несколько университетов, придется хорошенько потрудиться. Большинство заявлений подается в период с декабря по февраль, а ответ приходит примерно в феврале или марте. Если вас примут, то до апреля нужно решить, хотите ли вы учиться по этой программе. Получив положительный ответ, не пытайтесь выбрать «лучшее» — идите туда, где, на ваш взгляд, будете счастливы!

3.1.3. Заключение по академическому образованию

В целом академические программы по Data Science хорошо подходят людям, которые хотят получить всестороннее образование и могут себе это позволить. Можно перейти из другой сферы деятельности, где не приходилось заниматься программированием или технической работой, например из маркетинга. Такая программа позволит изучить все составляющие науки о данных в удобном ритме.

Академические программы *не* подходят людям, у которых уже есть многие из требуемых навыков: для них это будет слишком долго и дорого и в итоге не окупится. К тому же преподаватели не работают непосредственно в отрасли, и то небольшое количество новых знаний, которое они дадут, может вовсе не пригодиться на практике. Возможно, вам придется получить практический опыт на стажировке во время программы, чтобы дополнить свое образование.

Если вы считаете, что вам нужно углубленное обучение, чтобы стать дата-сайентистом, то вперед. Начинайте искать подходящее учебное заведение. Если же вы чувствуете, что такое обучение потребует слишком больших затрат и вместо этого есть более легкий путь, рассмотрите варианты, описанные в следующих разделах.

Нужна ли мне кандидатская степень для работы в Data Science?

Скорее всего, нет.

Кандидат наук — это степень, на получение которой уходит много лет и которая готовит обучающихся к должности профессора. Придется потратить годы на исследования, чтобы найти новый метод, который не особо лучше предыдущего. Нужно публиковаться в академических журналах и продвигать новейшие исследования в конкретной области. Но, как мы уже говорили в главах 1 и 2, та небольшая работа, которую выполняет дата-сайентист, похожа на научное исследование. Специалист по работе с данными гораздо меньше заботится о поиске элегантного искусного решения — ему достаточно чего-то работающего.

Небольшое число вакансий в области Data Science требует наличия степени кандидата наук. Но навыки, приобретенные в аспирантуре, редко бывают необходимыми для работы; как правило, такое требование говорит о статусе должности. Знаний, которые вы получите на программах магистратуры или бакалавриата, будет достаточно для большинства должностей в этой области.

Кроме того, у степени кандидата наук высокая цена, и речь не только о деньгах. Подумайте только, что за те семь лет, которые уйдут на ее получение, вы могли бы проработать в компании, улучшить свои навыки и зарабатывать гораздо больше денег.

Конечно, можно пойти и получить кандидатскую степень, а затем стать дата-сайентистом, но не позволяйте никому говорить, что вам без этого никак.

3.2. Буткемпы

Буткемпы — это интенсивные курсы продолжительностью 8–15 недель, которые организовывают тренинговые компании вроде Metis и Galvanize в США. Каждый день на лекции от специалистов отрасли, практику и работу над проектами уходит около восьми часов. В конце слушатели обычно выступают с финальным проектом перед аудиторией из сотрудников компаний, которым нужны дата-сайентисты. В идеале далее происходят собеседование и прием на работу.

Буткемпы дают много знаний за очень короткий срок, а это значит, что они подойдут тем, у кого уже есть большинство нужных навыков, но нескольких все же не хватает. Представьте себе нейробиолога, которому по работе приходилось заниматься программированием. На курсе по анализу данных он может пройти темы вроде логистической регрессии и баз данных SQL. С этими навыками и опытом работы в науке такой человек будет готов к работе в Data Science. Иногда

лучшее в буткемпе — не сами знания, а уверенность, которую дает программа, что вы действительно можете выполнять работу в DS.

3.2.1. Чему можно научиться

У хорошего буткемпа есть оптимизированная программа, которая научит вас всему необходимому для устройства на работу в Data Science, и не более того. Она дает не только технические навыки, но и возможность работать над проектами и общаться с людьми. В следующих разделах подробно описано, чего вам следует ожидать от буткемпа.

НАВЫКИ

Буткемп — это отличное дополнение к имеющемуся образованию. После него вы сможете быстро получить работу без потери двух лет на учебу (например, в магистратуре). Это может быть особенно приятно, если у вас уже есть степень магистра в области, не связанной с Data Science. В буткемпе вы обычно получаете следующие навыки:

- *Введение в статистику*. Этот курс включает методы прогнозирования на основе данных, например линейную и логистическую регрессию, а также методы тестирования, которые вы можете использовать в работе, например t-критерий Стьюдента. Из-за очень ограниченного времени вы не успеете разобраться в том, почему эти методы работают, но зато много узнаете об их применении.
- *Методы машинного обучения*. В программе расскажут об алгоритмах МО, таких как случайные леса и метод опорных векторов, и научат пользоваться ими с помощью разделения данных на тренировочные и испытательные наборы и выполнения перекрестной проверки. Можно изучать алгоритмы для конкретных случаев, например для обработки естественного языка или для поисковиков. Если вы не поняли ни слова из этого абзаца, возможно, буткемп для вас самое то!
- *Программирование на R или Python* (*средний уровень*). Вы изучите основы хранения и работы с данными во фреймах — как их складывать, фильтровать и строить графики. Научитесь использовать методы статистики и МО в выбранной программе. Скорее всего, вы будете проходить только один из этих языков, так что второй придется освоить самостоятельно, если вдруг он понадобится вам для работы.
- *Реальные кейсы*. Вы не только изучите разные алгоритмы, но и узнаете, как их можно применять на практике. Например, как с помощью логической регрессии спрогнозировать, когда клиент откажется от подписки на продукт, или как использовать алгоритм кластеризации для сегментирования покупателей в маркетинговых целях. Эти знания чрезвычайно полезны для устройства на работу, и на собеседованиях можно часто услышать вопросы на эту тему.

ПРОЕКТЫ

В буткемпах много внимания уделяется проектам. Вместо того чтобы слушать лекции по восемь часов в день, бóльшую часть времени вы будете работать над проектами, которые помогут вам лучше понять Data Science и начать работу над собственным портфолио (тема главы 4). Это огромный плюс по сравнению с академическим образованием, потому что эти навыки больше пригодятся в компаниях, где обязанности часто напоминают работу над проектами.

Для проекта сначала нужно собрать данные. Для этого можно использовать веб-API, созданный компанией для извлечения своих данных, скрейпить веб-сайты или же взять существующие публичные датасеты с таких ресурсов, как государственные веб-сайты. Затем вы будете загружать данные в R или Python, писать сценарии для управления и запускать на них модели МО. Полученные результаты понадобятся для презентации или отчета.

Для выполнения всего этого вовсе не обязательно идти в буткемп. Вообще-то, глава 4 этой книги полностью посвящена тому, как можно самостоятельно выполнять DS-проекты. Зато в буткемпе есть преподаватели, которые будут направлять вас и помогать с проектом, если что-то пойдет не так. Трудно сохранять мотивацию, если вы работаете в одиночку, и легко застрять с чем-нибудь, когда рядом нет человека, к которому можно обратиться за помощью.

СЕТЬ

Многие люди после буткемпов строят успешную карьеру в таких компаниях, как Google и Facebook. В подобные организации можно проскочить через сообщества выпускников. В буткемпы порой приглашают DS-спикеров, а на защиту вашей итоговой работы могут прийти представители компаний. Связи с такими людьми помогают устроиться в их организации. Эту особенность стоит выделить, ведь когда дело доходит до поиска работы, ворота в компанию с вакансиями в Data Science могут решить дело.

Помимо знакомства с людьми в процессе обучения можно использовать такие инструменты, как LinkedIn, для связи с выпускниками буткемпов. Эти люди могут помочь с устройством в компанию, где они работают, или хотя бы подсказать, как выбрать подходящую.

Все эти варианты предполагают, что вы должны действовать активно, например общаться со спикерами после презентаций или писать в социальных сетях людям, с которыми вы прежде не общались. Это может пугать, особенно если вам не слишком-то комфортно болтать с незнакомцами, но именно так вы получите максимум пользы от курсов. Ознакомьтесь с главой 6, чтобы узнать, как написать эффективный запрос на нетворкинг.

3.2.2. Цена

По сравнению с самообразованием у буткемпов есть один серьезный недостаток — цена, которая обычно составляет от $15 000 до $20 000. Хотя есть вариант получить стипендию, которая покроет часть стоимости обучения, нужно еще помнить об издержках, связанных с невозможностью работать полный день (а порой и неполный) во время программы. Кроме того, после буткемпа, скорее всего, придется еще несколько месяцев искать работу. Во время учебы делать это не получится из-за нехватки времени и навыков в Data Science, и, даже если вас примут на должность, весь процесс может занять несколько месяцев от момента отклика до первого дня работы. Короче говоря, буткемп может оставить вас безработным на шесть, а то и девять месяцев. Если у вас есть возможность самостоятельно изучать Data Science в свободное время или учиться на работе, то можно продолжать зарабатывать и не платить за обучение и таким образом сэкономить десятки тысяч долларов.

3.2.3. Выбор программы

Количество вариантов буткемпов зависит от того, где вы живете. Если вы хотите посещать их очно, то, вероятнее всего, даже в большом городе выбор доступных программ будет небольшим. А из провинции и вовсе придется на время перебираться в город, что увеличит стоимость программы и серьезно изменит вашу жизнь.

С другой стороны, есть онлайн-версии буткемпов по Data Science. Однако учтите: как и в случае с магистратурой, один из плюсов очных буткемпов в том, что люди вокруг будут мотивировать вас и помогать сосредоточиться на обучении. Если вы выберете онлайн-формат, то лишитесь этого преимущества и ваш буткемп за $20 000 может сравняться по эффективности с дешевыми или вовсе бесплатными открытыми онлайн-курсами.

При выборе буткемпа в вашем регионе не забудьте посетить аудитории, пообщаться с несколькими преподавателями и оценить, где вам комфортнее. Но будьте осторожны: везде хватает людей, которые стремятся выкачать деньги из тех, кто пытается стать дата-сайентистом. Если вы не будете осмотрительны, то можете попасть на программу, которая оставит вас без работы, зато с долгом в десятки тысяч долларов. При выборе буткемпа крайне важно пообщаться с теми, кто его окончил. Есть ли успешные выпускники этой программы на LinkedIn? Если да, пообщайтесь с ними и спросите, как они оценивают полученный опыт. Если вы не найдете таких людей на LinkedIn, это должно стать тревожным звоночком.

3.2.4. Заключение по DS-буткемпам

Буткемпы могут отлично подойти тем, кто хочет сменить профессию и уже немного разбирается в основах Data Science. Они также могут быть полезны для тех, кто только оканчивает университет и хочет пополнить портфолио DS-проектами,

прежде чем искать работу. Тем не менее буткемпы не рассчитаны на то, чтобы прокачать вас с нуля до 60-го уровня; у большинства из них есть высокие требования к зачислению, и нужно иметь опыт работы с основами статистики и программирования, чтобы поступить, а затем получить максимум от программы.

3.3. Работа с Data Science в вашей компании

Вы можете оказаться в сфере, связанной с Data Science. Необычный, но часто очень эффективный способ освоить это направление — начать все больше работать с данными в рамках текущей должности. Возможно, вы бизнесмен, который заставляет DS-отчеты звучать по-деловому. Тогда попробуйте добавлять в них собственные графики. Или, может быть, вы как работник финансовой сферы составляете электронные таблицы — их можно перенести на R или Python.

Рассмотрим гипотетическую Эмбер, человека, который несколько лет проработал в отделе маркетинговых исследований, проводил опросы клиентов и использовал графический интерфейс пользователя (GUI) для сбора результатов опроса. У Эмбер есть опыт работы в социологии и немного навыков программирования, полученных за время учебы. Она часто работает с отделом анализа данных, которому передает результаты опросов и объясняет их суть для использования в моделях. Со временем Эмбер начинает выполнять небольшую работу для DS-команды: извлекает функции в R, занимается визуализацией. Вскоре команда все больше и больше полагается на Эмбер. За это время ее навыки программирования и обработки данных действительно улучшаются. Через год она становится членом команды и работает на полную ставку, оставив сферу маркетинга в прошлом.

Попытка перейти к Data Science на текущем месте работы — отличный вариант, ведь так вы практически ничем не рискуете, зато очень мотивированы. Не нужно бросать работу ради дорогостоящего буткемпа или высшего образования; вы просто занимаетесь данными там, где это возможно. При этом такой подход мотивирует, потому что результаты вашей работы нужны остальным сотрудникам. Со временем вы сможете все больше заниматься Data Science, пока наконец это не станет вашей основной работой. Это совсем не похоже на вариант, когда вы сначала учитесь, а затем внезапно меняете сферу деятельности.

У Эмбер — бывшего маркетолога, а ныне дата-сайентиста, было кое-что еще:

- У нее были налаженные взаимоотношения с DS-отделом, который курировал ее работу.
- Она освоила основы программирования и визуализации данных.
- Она была достаточно мотивированной, чтобы изучать методы Data Science на работе.

- Отдел анализа данных смог поручить Эмбер небольшие проекты, которые со временем становились масштабнее, что в итоге позволило ей стать дата-сайентистом.

Если вы хотите освоить Data Science на работе, поищите места, где делают небольшие DS-проекты, и людей, готовых с ними помочь. Такие простые задачи, как создание отчета или автоматизация существующего, могут многому научить.

Одно важное замечание для выбравших этот путь: никогда не напрягайте других. Это может быть очевидным, например если вы неоднократно просите очистить для вас наборы данных, или менее явным, скажем если вы постоянно просите кого-то проверить вашу работу. Вы также можете неумышленно нагрузить команду, добавив новые инструменты. Если вы из финансового отдела и все, кроме вас (теперь вы используете R), работают с Microsoft Excel, то вы все усложнили. Даже обращение к кому-то с просьбой дать вам задание может быть людям в тягость, потому что тогда придется думать, чем вас занять. Так что старайтесь не создавать проблемы другим людям.

Две точки зрения на диалог

Что вы говорите: «Я рад помочь чем могу — просто дайте мне знать как! Спасибо!»

Что, по вашему мнению, слышат: «Я человек, который хочет работать на вас. Вы можете передать мне этот увлекательный, но простой проект, который так долго ждал своего часа, и я сделаю его за вас!»

Что слышат на самом деле: «Привет! Я хочу быть полезным, но понятия не имею, что вам нужно. Я также не знаю, какие из моих навыков будут полезны, так что удачи вам в поиске задачи для меня. Кроме того, если вы каким-то образом найдете задачу, которая мне идеально подходит, вам еще придется пересмотреть ее несколько раз, прежде чем я смогу с ней работать. Все это отнимет у вас и без того недостающее время. Спасибо!»

Чтобы этот путь был эффективным, нужно использовать несколько ключевых стратегий:

- *Проявляйте инициативу*. Чем больше работы вы сможете выполнить до того, как вас об этом попросят, тем более независимым станете и тем меньше будете обременять команду. У группы дата-сайентистов может быть скучная задача, на которую уйдет много времени, например маркировка данных или создание простого отчета. Можно предложить им свою помощь. Но будьте осторожны с самостоятельностью: может так получиться, что вместо пользы она принесет команде одну только необходимость все переделывать. Однако если вы можете приступить к задаче, в которую остальные позже внесут свой вклад, возможно, вы сэкономите для команды много времени.
- *Осваивайте навыки по одному*, не хватайтесь за все сразу. Выделите один навык, который вы хотите изучить в процессе работы, и принимайтесь за дело.

Например, можно научиться составлять отчеты с помощью R, потому что группа дата-сайентистов постоянно этим занимается. Взявшись за небольшой проект для помощи команде, можно добавить новый навык в свой арсенал. После этого можно приступать к изучению следующего.

- *Четко сформулируйте свои намерения*. Довольно быстро все поймут, что вы берете дополнительную работу для перехода в Data Science. Если вы проявите инициативу и дадите DS-команде понять, что хотите научиться большему, она может придумать, как вам помочь. Кроме того, члены команды будут учитывать вашу неопытность, потому что они когда-то тоже были новичками.
- *Не будьте чересчур напористым*. Помочь человеку стать дата-сайентистом — это огромный труд, а команды уже и так перегружены работой. Если вдруг окажется, что ни у кого нет времени или возможности помочь вам, не принимайте это на свой счет. Иногда напоминать о себе — нормально, но, если вы будете слишком настойчивы, команда быстро почувствует себя некомфортно. Участники будут рассматривать вас не как потенциальный ресурс, а как источник неудобства.

Когда возможностей нет

Вы можете оказаться в ситуации, когда на вашей нынешней должности нет возможности заниматься данными. Возможно, какие-то рабочие ограничения не позволяют вам использовать R или Python или реализовать методы анализа данных. В таком случае вам, возможно, придется принять решительные меры. Уйти с работы, чтобы пойти в буткемп или учиться в институте, — рискованный, но эффективный шаг, который поможет перейти на новый уровень. В свободное время также можно учиться самостоятельно, но у этого метода есть масса недостатков (см. раздел 3.4). Другой вариант — попытаться найти другую работу в своей области, которая открывала бы перспективы узнать больше. Но никто не гарантирует, что на новом месте вы получите обещанное.

Среди этих вариантов нет легких путей, но, к сожалению, такова жизнь. Чтобы получить работу в Data Science, придется потрудиться, но оно того стоит.

3.3.1. Выводы об обучении на работе

Обучение без отрыва от производства может быть эффективным способом стать дата-сайентистом при условии, что на вашей работе можно применить навыки в области анализа данных и есть люди, которые могут вас наставлять. При соблюдении этих условий такой путь — отличный вариант, но так бывает далеко не всегда. Если вы считаете, что он вам подходит, мы настоятельно рекомендуем выбрать именно его. Работа не всегда позволяет учиться без отрыва от производства, так что воспользуйтесь возможностью, если она есть.

3.4. Самообучение

Data Science посвящено огромное количество книг (например, эта) и множество онлайн-курсов. Они обещают научить вас основам DS, а также углубленным техническим навыкам (и по разумной цене) на практике. Эти курсы и книги, а также все блоги по Data Science, учебные пособия и ответы на Stack Overflow могут дать неплохую базу.

Такие материалы отлично подходят для приобретения отдельных навыков. Например, если вы хотите разобраться в глубоком обучении, книга может стать отличным помощником. А для изучения основ R и Python можно для начала пройти онлайн-курс.

Самостоятельно изучать Data Science по книгам и онлайн-курсам — это все равно что учиться играть на музыкальном инструменте по видео на YouTube или изучать что-либо еще без преподавателя: ценность такого подхода в основном зависит от вашей настойчивости. На освоение навыков с нуля могут уйти сотни, а то и тысячи часов. И правда, как можно сосредоточиться на Data Science, когда на соседней вкладке открыты лучшие подборки TikTok? Также трудно понять, с чего начать. Если вы хотите изучить все темы, то кто подскажет, какую книгу прочесть первой (может быть, эту)?

Самообучение означает, что у вас нет преподавателя или примера, на который стоит равняться. Без учителя, которому можно задавать вопросы, как в буткемпе или университете, вы не узнаете, все ли делаете правильно и что нужно делать дальше. Время будет потрачено впустую, если у вас нет четкого направления или вы вообще выбрали неправильный путь. Лучший способ восполнить отсутствие преподавателя — найти сообщество людей, в котором можно задавать вопросы. Отличный пример — программа TidyTuesday (https://github.com/rfordatascience/tidytuesday), инициированная Томасом Моком; каждый вторник начинающие и практикующие дата-сайентисты используют R для решения DS-задач.

Если вы все же решите учиться самостоятельно, важен грамотный подход. Книги и видео — это здорово, но вы узнаете гораздо больше, если будете практиковаться и делать выводы на основании проделанной работы. Другими словами, чтение о велосипедах может быть познавательным, но вы никогда не научитесь кататься, не сев на велосипед. Обязательно найдите проект, которым хотите заниматься, например найти в наборе данных интересные результаты, создать модель машинного обучения и API или использовать нейронную сеть для генерации текста. В главе 4 мы рассмотрим такие проекты подробнее. При других способах изучения Data Science проекты могут пригодиться для создания портфолио, но, когда вы учитесь самостоятельно, проекты играют именно образовательную роль.

3.4.1. Выводы о самообучении

Учиться самостоятельно сложно, но можно. Вы должны уметь определять учебный план, сохранять достаточную мотивацию и делать все это без наставника или преподавателя, который мог бы вам помочь. Вам также будет труднее продемонстрировать свою квалификацию в резюме, чем в других случаях. Из всех предложенных нами способов стать дата-сайентистом этому мы отдаем наименьшее предпочтение, поскольку при самообучении многое может пойти не так. Кроме того, многим попросту не удается сохранять сосредоточенность. Если вам нужно освоить какой-то один навык или технологию, такой способ может подойти, но для того, чтобы изучить все, что нужно знать специалисту, лучше выбрать другой путь.

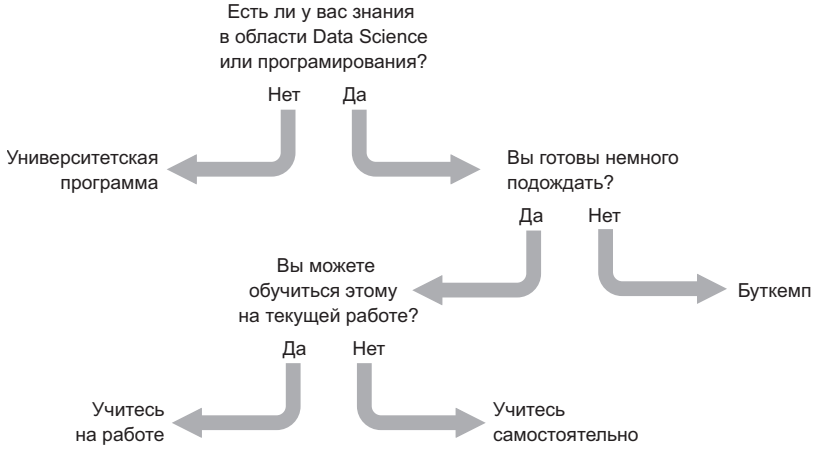


Рис. 3.1. Процесс принятия решения об изучении Data Science

3.5. Как сделать выбор

Как выбрать один из этих четырех совершенно разных подходов? Все принимают решения по-разному, но мы предлагаем ответить на три вопроса (рис. 3.1):

1. *Есть ли у вас знания в области Data Science*? В частности, знаете ли вы хоть один язык программирования, не учитывая легкие курсовые работы? Умеете запрашивать данные из базы SQL? Понимаете, что такое линейная регрессия?
   а) Если ваш ответ: «Нет, мне нужно многому научиться», вам, вероятно, лучше всего подойдет магистратура. Во время обучения на этой программе вы будете изучать различные темы в течение достаточно длительного времени и сможете хорошо их освоить.
   б) Если ваш ответ: «Да, я это знаю», переходите к вопросу 2.

2. *Согласны ли вы потратить год или даже больше на приобретение навыков, вместо того чтобы просидеть без работы всего 6–9 месяцев и стать дата-сайентистом быстрее*? Сложно быстро освоить навыки с нуля, если вы сосредоточены исключительно на обучении; с полноценной работой сделать это будет еще сложнее. Вы готовы потратить больше времени, чтобы сохранить занятость на полный рабочий день?
   а) Если ваш ответ: «Нет, мне нужно действовать быстро», запишитесь на курсы. Через три месяца вы освоите тонну информации и будете готовы приступить к поиску новой работы, что может дополнительно занять от трех месяцев до полугода.
   б) Если ваш ответ: «Да, не хочу торопиться», переходите к вопросу 3.
3. *Можете ли вы изучать Data Science на своей нынешней работе*? Можете ли вы делать всякие вещи с данными на текущей должности, например проводить анализ, хранить что-то в SQL или попробовать R или Python? Есть ли команда, которая могла бы наставлять вас или давать небольшие задания?
   а) Если ваш ответ: «Да, я могу здесь учиться», тогда действуйте и используйте свою работу как трамплин в Data Science.
   б) Если вы ответили: «Нет, у меня нет такой возможности», пора переходить к книгам и онлайн-курсам.

Эти вопросы должны стать вашей отправной точкой, однако вам не нужно принимать единственное окончательное решение. Вы можете начать с книг, а затем пойти в буткемп, если захотите двигаться быстрее. Также можно пойти в вечернюю магистратуру и заниматься Data Science на текущей работе. Однозначно правильного ответа нет; важно найти решение, которое подходит именно вам. Если что-то идет не так, меняйте методы до тех пор, пока не подберете работающий.

Выбрав маршрут, следуйте ему! Поступите наконец в магистратуру, запишитесь в буткемп или купите нужные книги и начните читать. В целях этой книги будем считать, что уже прошло какое-то время и вы успешно освоили навыки, необходимые дата-сайентисту. В следующих нескольких главах они пригодятся для портфолио, благодаря которому вы сможете получить первую работу в Data Science.

3.6. Интервью с Джулией Силдж, дата-сайентистом и инженером-программистом RStudio

Джулия Силдж (Julia Silge) известна благодаря своему блогу о Data Science, а также разработанному ею и Дэвидом Робинсоном (David Robinson) пакетом tidytext, который является краеугольным камнем естественной обработки языка в R и был скачан более 700 000 раз. Они также совместно написали книгу *Text Mining with R:*

A Tidy Approach (O'Reilly). Джулия несколько лет работала дата-сайентистом в Stack Overflow, а сейчас разрабатывает инструменты машинного обучения с открытым исходным кодом в RStudio.

Прежде чем стать дата-сайентистом, вы работали в академической сфере; как полученные навыки помогли вам в нынешней профессии?

Занимаясь исследованиями, я иногда собирала реальные данные. Этот опыт научил меня думать о процессах их создания. В том случае это был результат физического процесса, к которому я могла прикоснуться. Я на самом деле могла видеть, почему данные были неупорядоченными или почему мы не получили определенный результат в конкретной ситуации. Я вижу прямую взаимосвязь с работой, которой я семь лет занималась в технологической компании, чья деятельность связана с веб-данными. Там был какой-то процесс, который их генерировал, а я должна была тщательно следить за записью и правильностью его выполнения. Этот опыт работы с реальными данными определяет мой подход к разработке инструментов машинного обучения.

Еще до того, как стать дата-сайентистом, я научилась общаться и обучать. Я преподавала в колледже несколько лет, а еще работала на местах, предполагающих общение с клиентами. Таким образом, я развивала свое представление об определенной сфере и пыталась передать эти знания другому человеку. Я твердо верю, что это часть обязанностей большинства специалистов в Data Science. Если просто обучить какую-то модель или провести статистический анализ, это будет не настолько ценно, как если взять ту же модель или анализ и объяснить, что все это означает, как оно работает или как это можно реализовать в более широком контексте.

Где вы приобретали необходимые навыки для работы в Data Science?

Конечно, я думаю, что образовательные программы, буткемпы и онлайн-материалы — отличные варианты для разных людей в разных ситуациях. Учитывая, что у меня уже была кандидатская степень, мне не хотелось возвращаться в университет и тратить еще больше денег. Признаюсь, я подала заявку в пару буткемпов, а они меня прокатили! Когда я решила сменить карьеру, то понимала, что справлюсь с этой работой, но мне предстояло убедить в этом других. Мне также нужно было освежить знания о машинном обучении и некоторых методах, потому что, когда я училась в аспирантуре, современное МО еще не дошло до астрофизики.

Я выбрала путь онлайн-курсов и стала заниматься самообразованием. Иногда я в шутку говорю, что прошла все существующие MOOC (массовые открытые дистанционные курсы): их было действительно *много*. Я взяла перерыв на полгода в месте, где работала на полставки, и бросила все силы на курсы. Я давно не училась и была в восторге от материала. Какое-то время я не занималась анализом данных, поэтому вернуться к нему было действительно здорово!

Выбирая карьеру в Data Science, вы знали, чем конкретно хотели бы заниматься?

Рассматривая разные варианты будущего, я слышала о том, как специалисты рассуждали о различиях между *анализом* и *построением* в Data Science, и видела себя стопроцентным аналитиком. Я хотела быть не столько инженером, сколько ученым — человеком, который стремится понимать суть и отвечать на вопросы, а не заниматься построением. Так началась моя карьера. Бóльшую часть времени я была единственным дата-сайентистом в Stack Overflow и работала в команде с очень талантливыми дата-инженерами, настоящими мастерами своего дела. Как единственный специалист по данным я занималась анализом и построением моделей. Теперь, когда я работаю над инструментами с открытым исходным кодом, моя должность называется «инженер ПО» и я трачу больше сил на построение, чем на анализ.

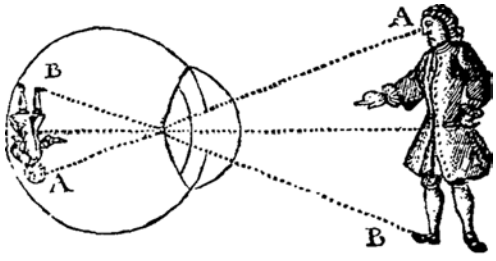
Что бы вы посоветовали тем, кто хочет получить навыки, необходимые дата-сайентисту?

Я бы хотела особенно подчеркнуть вот что: вы должны продемонстрировать, что справитесь с этой работой. Это можно сделать по-разному в зависимости от ситуации. Data Science все еще остается относительно молодой сферой, и многие до конца не уверены в том, что значит быть специалистом по данным и кто может им стать. Границы обязанностей по-прежнему достаточно размыты, а поскольку эта должность хорошо оплачивается, компании сильно рискуют при найме таких сотрудников и, естественно, желают избежать ошибок. Организация должна быть уверена, что кандидат справится с работой. Я видела, как люди ради этого делали вклад в открытый исходный код, рассказывали о своих проектах на местных встречах и создавали портфолио в блоге или на GitHub. Я же прошла курсы MOOC, изучила то, что считала важным, и завела блог. Я надеялась, что все эти проекты и посты в блогах можно будет обсудить на собеседовании.

Итоги

- Четыре проверенных способа получить навыки для работы в Data Science: академические программы, буткемпы, практика на текущей работе и самостоятельное обучение.
- Каждый из этих способов имеет свои преимущества и недостатки с точки зрения материалов, времени и самоорганизации.
- Чтобы выбрать для себя подходящий способ, подумайте о своих сильных сторонах, готовых навыках и доступных ресурсах.

4



Создание портфолио

В этой главе

- Как создать интересный DS-проект.
- Как начать блог.
- Как провести проект от начала до конца.

Вы окончили буткемп, магистратуру, прошли онлайн-курсы или подготовили несколько проектов с данными на текущей работе. Поздравляем — вы готовы начать работу в Data Science! Так ведь?

Почти. Вторая часть этой книги посвящена тому, как найти вакансию, отправить отклик и устроиться дата-сайентистом. И вы, безусловно, можете начать прямо сейчас. Но есть еще одна вещь, которая действительно поможет добиться успеха, — портфолио. *Портфолио* — это набор DS-проектов, которые можно показать людям, чтобы они поняли, с какими задачами вы можете справляться.

Сильное портфолио состоит из двух основных частей: репозиториев GitHub (*Git-репозитории*) и блога. В Git-репозитории размещают код проекта, а в блоге можно похвастаться навыками коммуникации и остальной частью работы, не связанной с кодом. Большинство людей не захотят вникать в тысячи строк кода (ваш Git-репозиторий); им нужно получить быстрое объяснение того, что вы сделали и почему это важно (для этого нужен блог). И кто знает, возможно, дата-сайентисты со всего мира станут читать ваши статьи. Во второй части этой главы мы расскажем, что в блоге стоит писать не только о проведенном анализе или построенных моделях; можно объяснять статистические методы, написать

учебное пособие по анализу текста или даже поделиться секретами карьерного роста (например, на тему, как вы выбрали программу обучения).

Это не означает, что для успеха обязательно нужен блог или Git-репозиторий. На самом деле у большинства специалистов всего этого нет, а люди постоянно находят работу и без портфолио. Но создание портфолио — отличный способ выделиться и попрактиковаться в работе с данными. Надеемся, вам понравится!

В этой главе мы расскажем об этапах создания хорошего портфолио. Первая часть посвящена созданию DS-проекта и его систематизации на GitHub. Во второй части описываются наиболее эффективные методы создания и ведения блога. Дальше мы расскажем о двух наших реальных проектах, чтобы вы могли воочию увидеть и понять процесс от начала до конца.

4.1. Создание проекта

DS-проект начинается с двух вещей: интересного набора данных и вопроса. К примеру, можно взять данные государственной переписи населения и спросить: «Как со временем меняются демографические данные по стране?» Сочетание вопроса и данных составляет ядро проекта (рис. 4.1). С этими двумя параметрами на руках можно приступать к анализу.

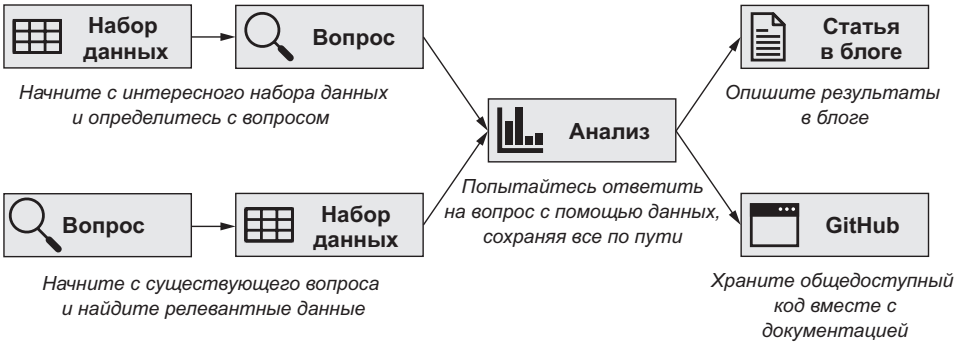


Рис. 4.1. Последовательность создания DS-проекта

4.1.1. Найдите данные и задайте вопрос

Размышляя над тем, какие данные использовать, самое важное — чтобы они были интересными. Почему вы хотите использовать именно их? Выбор данных — это способ проявить свою личность или знания в предметной области, которые вы получили в ходе предыдущей карьеры или учебы. Например, если вы увлекаетесь стилем, присмотритесь к статьям о «Неделе моды» и узнайте, какие изменения произошли за последние 20 лет. Если вы отлично бегаете, можно показать, как

со временем менялись ваши результаты, и, возможно, проверить, зависит ли скорость от погоды.

А вот чего делать не следует, так это использовать данные о «Титанике», БД MNIST или любые другие популярные наборы. Дело не в том, что они чем-то плохи; вероятнее всего, вы не найдете ничего нового, что могло бы удивить и заинтриговать работодателей или дать им больше представления о вас.

Иногда к датасету приводит сам вопрос. Например, вам может быть любопытно, как изменился гендерный состав студентов по разным специальностям в колледжах и связано ли это изменение со средним заработком после выпуска. Затем вы гуглите в поисках наиболее подходящего источника этих данных.

Но, возможно, у вас нет вопроса, на который вы давно мечтали ответить с помощью навыков в Data Science. В таком случае можно попробовать сначала поискать датасеты и посмотреть, получится ли задать к ним интересный вопрос. Вот несколько советов, с чего начать:

- Kaggle.com — начинался как веб-сайт для соревнований в Data Science. Компании публикуют датасеты и вопрос, предлагая приз за лучший ответ. Вопросы связаны с моделями МО для прогнозирования: например, сможет ли человек выплатить кредит или за сколько будет продан дом. Так что пользователи могут сравнивать модели на основе их производительности на тестовой группе выборки и получить показатели производительности для каждой из этих моделей. У Kaggle также есть форумы и «ядра», где участники делятся своим кодом. В результате у Kaggle есть тысячи датасетов с сопутствующими вопросами и примерами того, как их анализировали другие люди.
- В самом большом преимуществе Kaggle кроется и его главный недостаток: предоставив вам (как правило, очищенный) датасет и задачу, он сделает бóльшую часть работы за вас. Тысячи людей решают одно и то же, поэтому сложно предложить выдающееся решение. Один из способов использовать Kaggle — взять имеющийся датасет, но задать другой вопрос или провести эксплораторный факторный анализ. Но в целом мы думаем, что с Kaggle лучше всего учиться, когда вы решаете задачу, а затем смотрите, как это сделали остальные. Этот ресурс подойдет для того, чтобы понять, как работают чужие модели, а не для пополнения портфолио.
- *Датасеты из новостей*. В последнее время многие новостные компании начали публиковать свои данные. Например, FiveThirtyEight.com — сайт, посвященный анализу опросов общественного мнения, политике, экономике и спортивным блогам, — хранит данные, которые используются в статьях, и даже ссылается на сырые данные в своих публикациях. Хотя эти датасеты часто нужно чистить вручную, тот факт, что они появляются в новостях, означает, что к ним относятся очевидные вопросы.

- *API-интерфейсы*. API-интерфейсы (интерфейсы прикладного программирования) — это инструменты разработчика, которые позволяют получать доступ к данным напрямую от компаний. Вы знаете, как ввести URL-адрес и перейти на веб-сайт? API похожи на URL-адреса, но вместо веб-сайта вы получаете данные. Некоторые компании с полезными API-интерфейсами вроде The New York Times и Yelp позволяют вам получать их статьи и обзоры. У некоторых API даже есть пакеты R или Python, которые упрощают работу с ними. Например, rtweet для R позволяет быстро извлекать данные из Twitter, чтобы можно было найти, например, твиты с определенным хештегом, актуальные темы в Киото или любимые твиты Стивена Кинга. Учтите, что есть определенные ограничения и условия использования этих API. Например, Yelp ограничивает ежедневное количество вызовов до 5000, поэтому получить все просмотры не получится. API-интерфейсы отлично подходят для получения надежных, систематизированных данных из многих источников.
- *Открытые государственные данные*. Многие из них доступны в интернете. Можно найти информацию о переписи населения, занятости, опросах общественного мнения, а также множество данных местной администрации, например звонки в службы спасения Нью-Йорка или показатели дорожного движения. Иногда получается загрузить эти данные непосредственно в виде файла CSV; в других случаях понадобится API. На основании Закона о свободном доступе информации[1] вы даже можете отправлять запросы в государственные учреждения, чтобы получить данные, которых нет в открытом доступе. Государственная информация хороша тем, что она часто исчерпывающая и касается необычных вопросов, как, например, сведения о кличках каждого зарегистрированного животного в Сиэтле. К недостаткам такой информации можно отнести то, что она часто плохо отформатирована, например таблицы могут храниться в файлах PDF.
- *Ваши собственные данные*. Есть много ресурсов, где вы можете скачать данные о себе, два самых крупных из которых — это социальные сети и электронная почта. Но если вы используете приложения для отслеживания своей физической активности, списков книг, бюджета, сна и тому подобного, можно получить информацию и оттуда. Возможно, у вас получится создать чат-бота на основе переписки с супругом. Или определить наиболее часто употребляемые вами слова в твитах и отследить, как они менялись со временем. А может, стоит отслеживать потребление кофеина и заниматься спортом в течение месяца, чтобы попробовать спрогнозировать продолжительность и качество своего

[1] Freedom of Information Act, один из законов США. — *Примеч. ред.*

сна. Преимущество использования собственных данных в том, что проект будет стопроцентно уникален: никто ранее не видел эти данные!

- *Веб-скрейпинг.* Скрейпинг — это способ извлекать данные с веб-сайтов, не имеющих API, в основном через автоматический просмотр страниц и копирование. Можно написать программу, которая будет искать киносайты по списку из 100 актеров, скачивать их профили и списки фильмов, в которых они снимаются, и помещать эти данные в таблицу. Однако следует учитывать, что скрейпинг может быть запрещен правилами пользования сайтом и вас могут забанить, поэтому проверьте файл robots.txt. Будьте осторожны с сайтами: слишком большое количество запросов за короткое время может их обрушить. В то же время соблюдение правил и увеличение промежутка между запросами может сделать скрейпинг эффективным инструментом для получения уникальных датасетов.

Что делает сторонний проект интересным? Мы рекомендуем делать эксплораторный факторный анализ, в котором любой результат может быть чем-то полезен читателю или продемонстрирует ваши навыки. Можно создать интерактивную карту звонков на номер 311 в Сиэтле, размеченную цветом по категориям. Эта карта наглядно продемонстрирует ваши навыки визуализации и покажет, что вы можете описывать закономерности. С другой стороны, прогнозировать фондовый рынок, скорее всего, не получится, а с плохими результатами работодателю будет сложно оценить вас.

Еще один совет: загуглите свой вопрос и посмотрите на результаты. Если первыми в списке идут газетные статьи или сообщения в блогах, которые дают точный ответ на него, возможно, вам стоит пересмотреть свой подход. Иногда вы можете расширить чей-то анализ или улучшить его с помощью дополнительных данных, но для этого, возможно, придется начать все с нуля.

4.1.2. Выбор направления

Создание портфолио не должно отнимать много времени. Это именно та ситуация, где лучшее — враг хорошего. Сделать хоть что-то гораздо лучше, чем ничего. Первое и главное, что важно для работодателя, — убедиться, что вы умеете писать код и можете ответить на вопросы о данных. Возможно, вы переживаете, что люди посмеются над вашим кодом или скажут: «Хм, мы думали, что этот кандидат неплох, но гляньте на этот ужасный код!» Но такое вряд ли произойдет. Причина проста: ожидания работодателей зависят от опыта кандидата. Если вы начинающий специалист, от вас не ждут, что вы будете программировать как профессионал. Обычно в компании больше всего переживают, что вы вообще не умеете программировать.

Также неплохо поразмыслить над направлениями Data Science, которые мы рассмотрели в главе 1. Хотите специализироваться на визуализации? Сделайте интерактивный график с помощью D3. Хотите обрабатывать естественный язык? Используйте текстовые данные. Машинное обучение? Спрогнозируйте что-нибудь.

Заставьте себя выучить что-нибудь с помощью работы над проектом. Практика поможет вам заметить пробелы в знаниях. Например, если интересующие вас данные находятся на веб-сайтах, вы будете осваивать скрейпинг. Если вам не нравится, как выглядит какой-нибудь график, поработаете над улучшением визуализации. Выполнение проекта во время самообучения — хороший способ понять, куда двигаться дальше.

Распространенная проблема среди самостоятельно работающих специалистов — раздувание масштабов проекта. Такое происходит, когда дата-сайентист пытается сделать все сразу или продолжает добавлять все больше деталей по ходу проекта. Можно бесконечно улучшать/редактировать/дополнять его и так и не закончить работу. Вместо этого лучше думать, как в Голливуде, и создавать сиквелы. В проекте вы должны задать вопрос и ответить на него. А если вам кажется, что к нему еще надо будет вернуться, проект можно завершить вопросом или предложить тему для дальнейшего исследования (или даже написать «Продолжение следует…», если угодно).

Другая проблема — неспособность менять курс. Порой бывает, что нужные вам данные недоступны, либо их слишком мало или не получается очистить. Такие ситуации могут раздражать, и в эти моменты легко сдаться, но стоит все-таки постараться придумать, как спасти проект. Может быть, этой информации достаточно для того, чтобы написать туториал в блоге, например рассказать, как вы искали данные? Работодателям нужны люди, которые извлекают опыт из собственных ошибок и не боятся их признавать. Умение показать другим, что пошло не так, чтобы они могли избежать подобной ошибки, ценится по-прежнему.

4.1.3. Заполнение GitHub README

Возможно, вы уже работаете над собственными проектами в буткемпе или университете. Вы даже разместили свой код на GitHub. Этого достаточно?

Нет! Минимальное требование для полезного репозитория GitHub — заполнение файла README. Вам нужно ответить на несколько вопросов:

- *Что это за проект*? Какие данные в нем используются? На какой вопрос он отвечает? Что получилось: модель, система машинного обучения, информационная панель или отчет?
- *Как организован репозиторий*? Этот вопрос, конечно, подразумевает, что репозиторий действительно каким-то образом организован! Есть множество раз-

личных систем, но основная заключается в разделении вашего скрипта на части: получение (если необходимо) данных, их очистка, изучение и окончательный анализ. Таким образом, люди будут знать, как найти то, что им интересно. А для будущего работодателя это означает, что вы сможете хорошо организовывать информацию. Если при сдаче проекта вы отдадите сценарий на 5000 строк без единого комментария, никто не сможет его понять и использовать: компания не хочет так рисковать, нанимая кандидата. Умение управлять проектами также поможет вам в будущем: если вы захотите повторно использовать часть кода, то будете знать, где его искать.

Подготовить проект, сделать его общедоступным и задокументировать в репозитории GitHub — это хорошо, но по одному только коду очень сложно понять, что же в нем такого особенного. Когда вы приступаете к работе над проектом, вашим следующим шагом должен стать пост в блоге, где вы объясняете, почему то, что вы сделали, — круто и интересно. Никого не заинтересует pet_name_analysis.R, зато никто не пройдет мимо заголовка «Я использовал R, чтобы найти самые глупые клички питомцев!»

4.2. Создание блога

Благодаря блогу вы можете продемонстрировать свои проекты и размышления о них, кроме того, в блоге вы можете представить свою работу с нетехнической стороны. Знаем-знаем — вы только что научились всем этим замечательным техническим штукам и хотите похвастаться! Но работа в Data Science почти всегда влечет за собой передачу результатов непрофессиональной аудитории, а в блоге вы сможете попрактиковаться в переводе с технического на деловой язык.

4.2.1. Возможные темы

Предположим, вы завели блог. А будут ли ваши проекты вообще хоть кому-нибудь интересны? Вы еще толком не стали дата-сайентистом — чему у вас можно научиться?

Стоит запомнить одну вещь: лучше учить тех, кто отстает от вас всего на несколько шагов. Как только вы что-нибудь освоили, например научились использовать непрерывную интеграцию для пакета или создавать модели TensorFlow, у вас еще свежи воспоминания о своих ошибках и неудачах. Спустя годы трудно представить себе образ мышления новичка. Вы когда-нибудь занимались с очень компетентным преподавателем, который при этом совершенно не мог объяснить тему? Вы не сомневались в его профессионализме, но общение никак не складывалось, а преподаватель был крайне разочарован тем, что вы не схватывали все на лету.

Попробуйте разглядеть в читателе себя полгода назад. Что вы узнали с тех пор? Каких ресурсов вам не хватало? Этот опыт также отлично подходит для отслеживания собственного прогресса. В Data Science нужно столько всего выучить, что вы можете решить, будто делаете недостаточно; всегда полезно взять паузу и оценить свои достижения.

Вы можете разбить посты в блоге на четыре категории:

- *Туториалы, где много кода*. Они обучают, как что-то делать, например скрейпить веб-сайты или работать с глубоким обучением на Python. Вашими читателями, как правило, будут другие начинающие или практикующие дата-сайентисты. И хотя мы говорим «*с большим количеством кода*», текста в статье может быть едва ли не больше. Код сам себя не объяснит, так что вам придется описывать, что делает каждая его часть, почему вы решили сделать именно так и какие результаты получили.
- *Туториалы, где много теории*. Они могут объяснять статистические или математические концепции, например что такое эмпирический байесовский анализ или как работает анализ главных компонентов. В них также могут быть какие-нибудь уравнения или модели. Как и в предыдущем случае, вашей аудиторией будут в основном другие специалисты по данным, но вы должны писать так, чтобы вас мог понять любой, кто хоть немного разбирается в математике. Теоретические туториалы особенно хороши для демонстрации ваших коммуникативных навыков; есть стереотип, что многие технические специалисты, особенно с ученой степенью, не умеют объяснять понятия простым языком.
- *Ваш необычный проект*. Надеемся, в разделе 4.1 нам удалось убедить вас в том, что не стоит работать только над уникальным распознаванием медицинских изображений. Вы также можете проверить, в каких фильмах саги «Сумерки» использованы только слова из шекспировской «Бури». Например, Джулия Силдж использовала нейронные сети для генерации текста, который похож на произведения Джейн Остин. Посты в блоге могут описывать как процесс, так и результат в зависимости от того, какая часть проекта была наиболее интересной.
- *Делитесь своим опытом*. Не стоит публиковать в блоге только туториалы или заметки о DS-проектах. Можно поделиться своим опытом посещения встречи или конференции по Data Science: рассказать об интересных докладах, дать советы тем, кто собирается посетить их впервые, или поделиться некоторыми ресурсами от спикеров. Такие посты могут пригодиться людям, которые собираются посетить это мероприятие в следующем году либо не могут посещать конференции из-за нехватки денег или невозможности доехать. Опять же, этот тип публикаций позволяет потенциальным работодателям понять, как вы мыслите и общаетесь.

4.2.2. Выбор платформы

Но где же размещать статьи? У вас есть два основных варианта:

- *Создание собственного сайта*. Если вы работаете в R, мы предлагаем использовать пакет blogdown, который позволяет создать веб-сайт для блога с разметкой на R (круто, правда?). Если вы используете Python, то Hugo и Jekyll — это два варианта для создания статических веб-сайтов для блога. В них есть множество тем, созданных другими людьми, так что вы можете публиковать статьи с готовой разметкой. Не слишком заморачивайтесь насчет темы и стиля: просто выберите то, что вам понравится. Нет ничего хуже, чем перестать вести блог, увлекшись настройкой его оформления. Чем проще, тем лучше; постарайтесь выбрать тему, которая не надоест вам через полгода, ведь поменять ее может быть непросто.
- *Использование Medium или другой платформы для ведения блогов*. Medium — это бесплатная онлайн-платформа для публикаций. Сама компания обычно контент не создает, зато позволяет сотням тысяч авторов его размещать. Medium и подобные сайты — хороший вариант, если вам нужен быстрый старт, ведь с ним не придется беспокоиться о хостинге и создании веб-сайта: все, что от вас требуется, это нажать «New Post», написать и опубликовать статью. При этом вы даже можете получить больше трафика, когда люди будут искать на сайте такие термины, как *Data Science* или *Python*. Однако у этого варианта есть свой недостаток: вы зависите от платформы. Если, например, компания изменит бизнес-модель и сделает доступ к ресурсу платным, вы не сможете оставить свои публикации бесплатными. Вы также не сможете создать раздел с биографией или добавить другой непредусмотренный контент, например страницу со ссылками на ваши выступления.

Чаще всего люди хотят знать, какими должны быть объем и частота публикаций. Решайте сами. Мы видели, как авторы микроблогов публикуют короткие посты несколько раз в неделю. У других между публикациями проходят месяцы, но зато у них длинные статьи. Некоторые ограничения все же есть: вы ведь не хотите, чтобы ваши публикации напоминали роман Джеймса Джойса «Улисс»? Если ваш пост очень длинный, стоит разбить его на части. Ваша задача — показать, что вы умеете лаконично излагать информацию, поскольку это один из основных навыков в Data Science. Ни руководство компании, ни даже ваш непосредственный руководитель не хотят и, наверное, не должны слышать обо всех ваших неудачных попытках или двадцати испытанных подходах. Хотя неудачи и можно кратко описать, лучше поскорее перейти к сути и предложить окончательный способ решения. Из этого правила есть одно исключение: если ваш итоговый способ решения задачи удивит читателей. Если, например, вы не использовали самую популярную библиотеку, то можете объяснить, что сделали так, потому что оказалось, что она не работает.

Вас может беспокоить мысль, что никто не будет читать ваш блог и старания будут напрасны. Что ж, одна из главных причин вообще завести блог состоит в том, что он поможет вам при устройстве на работу. Вы сможете размещать в своем резюме ссылки на свои публикации при упоминании собственных DS-проектов и даже демонстрировать их на собеседованиях, особенно если там есть приятная интерактивная визуализация или информационные панели. Не обязательно иметь сотни или тысячи подписчиков. Да, было бы неплохо получить признание на Medium или быть упомянутым в информационном бюллетене компании, занимающейся Data Science, но гораздо важнее иметь аудиторию, которая будет читать, ценить и взаимодействовать с материалом.

Это не значит, что ничего нельзя сделать для увеличения количества читателей. Во-первых, вы должны рекламировать себя. Пусть это клише, но личный бренд помогает нетворкингу в долгосрочной перспективе. Даже если что-то кажется вам простым, это может быть неизвестно группе практикующих дата-сайентистов только потому, что область знаний очень велика. Ваши материалы могут читать даже сотрудники компаний, в которых вы хотите работать! Twitter — хорошее место для старта: вы можете поделиться новостью и использовать соответствующие хештеги, чтобы расширить круг читателей.

И все же ведение блога оправданно даже в том случае, если никто (кроме супруга и домашних животных) его не читает. Написание статей — полезный опыт, развивающий ваш навык структурирования мыслей. Как и преподавание, оно может помочь вам понять, что в чем-то вы не так сильны, как казалось.

4.3. Работа с примерами проектов

В этом разделе мы подробно рассмотрим два примера проектов: от появления идеи до проведения анализа и получения итогового результата. Для этого мы используем реальные проекты, принадлежащие авторам книги: создание веб-приложения для поиска наиболее подходящих вакансий для фрилансеров, занимающихся работой с данными, и обучение нейронных сетей на наборе данных о «запрещенных» автомобильных номерах.

4.3.1. Фрилансеры в Data Science

Эмили Робинсон

ВОПРОС

Будучи начинающим дата-сайентистом, я заинтересовалась фрилансом как одним из способов дополнительного заработка. *Фриланс* — это выполнение проектов для кого-то, у кого вы не работаете в штате, будь то человек или крупная компания.

Срок выполнения проектов может занимать от нескольких часов до нескольких месяцев полной занятости. Множество вакансий для фрилансеров можно найти на таких сайтах, как UpWork, но поскольку сфера Data Science очень обширна, вакансии в этой категории могут быть абсолютно разные, от веб-разработки до анализа в Excel и обработки терабайта данных на естественном языке. Я решила проверить, удастся ли мне помочь фрилансерам выбрать среди тысячи вакансий ту, которая им подойдет лучше всего.

АНАЛИЗ

Чтобы собрать данные, я использовала API сайта UpWork для получения открытых вакансий и профилей всех, кто находится в категории «Data Science и аналитика». В итоге у меня получилось 93 000 фрилансеров и 3000 вакансий. Хотя благодаря API доступ к данным стал относительно простым (поскольку не нужно было скрейпить), мне все равно пришлось создать функции для выполнения сотен вызовов API, делать обработку, когда эти вызовы не выполнялись, а затем преобразовать данные так, чтобы их можно было использовать. Зато так как данные не были готовы изначально, проект получился уникальным, а не как в случае с Kaggle, когда сотни людей работают над одним и тем же.

Получив хорошие, качественные данные, я провела эксплораторный анализ и изучила, как уровень образования и страна влияют на размер заработка фрилансеров. Кроме того, я составила график корреляции перечисленных фрилансерами навыков, на основании которого можно выделить различные группы: веб-разработчики (PHP, jQuery, HTML и CSS), специалисты по финансам и бухгалтерскому учету (финансовый учет, бухгалтерский учет и финансовый анализ), а также специалисты по сбору данных (ввод данных, лидогенерация, датамайнинг и веб-скрейпинг) наряду с «традиционным» набором навыков в Data Science (Python, машинное обучение, статистика и анализ данных).

Наконец, я придумала шкалу баллов для оценки сходства между профилем кандидата и описанием вакансии и объединила эту шкалу с совпадением навыков (как в резюме, так и в вакансии). В результате получился балльный счет для совпадения фрилансера и работы.

ГОТОВЫЙ ПРОДУКТ

Я не стала публиковать результаты в блоге, а вместо этого создала интерактивное веб-приложение, в котором можно было ввести текст своего профиля, указать навыки и требования к вакансии (например, минимальное количество отзывов о работодателе и время, которое займет работа). В результате доступные вакансии будут отфильтрованы в соответствии с этими требованиями и отсортированы по тому, насколько они подходят пользователю.

Я не позволила лучшему стать врагом хорошего; всегда есть много способов усовершенствовать проект. Я добавила вакансии всего один раз; теперь, спустя четыре года, ни одна из них больше не доступна, хотя приложение все еще работает. Чтобы оно было актуальным в долгосрочной перспективе, пришлось бы ежедневно подбирать вакансии и обновлять списки. Я также могла бы создать более сложный алгоритм сопоставления, ускорить начальное время загрузки приложения и сделать интерфейс привлекательнее. Но, несмотря на эти ограничения, я все же достигла нескольких важных целей. Проект показал, что я могу сделать что-то, с чем другие пользователи смогут взаимодействовать, а не ограничиваться статическим анализом, который хранится у меня на ноутбуке. Этот проект нашел свое применение в реальной жизни: благодаря ему фрилансеры могли найти работу. Наконец, я прошла через полный цикл DS-проекта: сбор данных, их очистку, выполнение эксплораторного анализа и получение результата.

4.3.2. Обучение нейронной сети на «неприличных» автомобильных номерах

Жаклин Нолис

ВОПРОС

По мере того как я росла как дата-сайентист, меня всегда удивляли радостные посты в блогах, где люди рассказывали о том, как обучали нейронные сети генерировать новые названия групп, покемонов и странные описания кулинарных рецептов. Я считала эти проекты великолепными, но не знала, как делать такое самостоятельно! Однажды я вспомнила, как слышала о датасете с автомобильными номерами, которые не приняли в Аризоне из-за их неблагозвучности. Будь у меня этот датасет, он бы идеально подошел, чтобы наконец научиться создавать нейронные сети, — я могла бы создавать собственные «неприличные» автомобильные номера (рис. 4.2)!

АНАЛИЗ

Отправив в Министерство транспорта штата Аризона запрос о предоставлении архивных документов, я получила список из тысяч «неприличных» автомобильных номеров. Тогда я ничего не знала о нейронных сетях, поэтому начала гуглить в поисках соответствующих постов. Будучи заядлым пользователем R, я была на седьмом небе от счастья, наткнувшись на пакет Keras от RStudio, который можно было использовать для создания нейронных сетей на R.

Я загрузила данные в R, а затем проверила пример пакета Keras от RStudio для генерации текста с помощью нейронных сетей. Я внесла изменения в код, чтобы

Рис. 4.2. Пример выходных данных нейронной сети генератора «неприличных» автомобильных номеров

обеспечить возможность работы с данными; пример RStudio был предназначен для генерации последовательностей длинного текста, но мне хотелось обучать модель на семизначных номерных знаках. Это означало создание нескольких точек датасета для обучения из каждого автомобильного знака (одна точка данных для прогнозирования каждого символа в номерном знаке).

Затем я обучила модель нейронной сети, хотя поначалу она не сработала. Я отложила проект и через месяц, вернувшись к нему, поняла, что данные обрабатываются неправильно. Когда я исправила эту проблему, результаты, созданные нейронной сетью, были фантастическими. В итоге, несмотря на мои незначительные изменения примера от RStudio, я продвинулась в создании и использовании нейронных сетей.

ГОТОВЫЙ ПРОДУКТ

Я рассказала о своем проекте в блоге, подробно описав, как я получила данные, как они обрабатываются для нейронной сети и как я изменила пример кода RStudio, чтобы он работал как положено. Пост был написан преимущественно в стиле: «Я в нейронных сетях новичок, но вот чему я научилась»; я не делала вид, будто уже знаю, как все это работает. В посте я разместила картинку, содержавшую выданный моей нейронной сетью текст и выглядевшую как номерной знак из Аризоны. Я также выложила код на GitHub. С тех пор как я написала этот пост и сделала код доступным, многие внесли в него свои изменения, чтобы создать собственные забавные нейронные сети. То, что я узнала, работая над этим нелепым проектом, в конечном итоге помогло мне создать высокоэффективные модели машинного обучения для важных консалтинговых услуг. Если изначальная работа кажется вам несерьезной, это не означает, что в ней нет никакой ценности!

4.4. Интервью с Дэвидом Робинсоном, дата-сайентистом

Дэвид Робинсон (David Robinson) является соавтором (с Джулией Силдж) пакета tidytext на R и книги *Text Mining with R*, выпущенной издательством O'Reilly. Также он написал электронную книгу *Introduction to Empirical Bayes: Examples from Baseball Statistics and the R packages broom and fuzzyjoin* и создал пакеты broom и fuzzyjoin для R. Дэвид окончил Принстонский университет со степенью кандидата наук в области количественной и вычислительной биологии. В своем популярном блоге varianceexplained.org Робинсон пишет о статистике, анализе данных, образовании и программировании на R.

Как вы начали вести блог?

Впервые я завел блог, когда начал искать работу перед получением кандидатской степени, так как осознал, что в интернете нет практически ничего, что демонстрировало бы мои навыки по программированию и статистике. Начав вести блог, я боялся, что напишу несколько постов, уже практически сформировавшихся в голове, и после этого мои идеи иссякнут. Однако я удивился, когда понял, что у меня постоянно появляются идеи для публикаций: датасеты, которые я собираюсь проанализировать, мысли, которыми мне хочется поделиться, методы, которым я хочу научить других. С тех пор я умеренно и систематически веду блог уже четыре года.

Есть ли какие-то особые возможности, которые вы получили благодаря онлайн-публикациям?

Свою первую работу я получил благодаря одному из оставленных мной сообщений в интернете. В Stack Overflow заинтересовались моим комментарием, который я оставил на их статистическом сайте, и они приняли меня на работу. Я оставил его несколько лет назад, но некоторые из инженеров компании заметили его и были впечатлены. Благодаря этому опыту я убедился в том, что публичная работа полезна. Иногда ее преимущества могут давать результат через месяцы или даже годы и дарить возможности, о которых нельзя было даже мечтать.

Как вы считаете, есть ли люди, которым публичная работа была бы особенно полезна?

Больше всего такая работа пригодится тем, чье резюме может недостаточно отражать навыки в Data Science, а также тем, у кого нет стандартной базы в виде диплома или опыта работы в сфере DS. Если у кандидата слабое резюме, мне сложно оценить, справится ли он со своими обязанностями. Мой любимый способ оценивать кандидата — просматривать анализ, который он проводил онлайн. Если у меня есть возможность увидеть созданные кем-то графики, прочесть, как

они создавались, как проводился разбор данных, я начинаю понимать, подходит ли человек на выбранную должность.

Как со временем изменилось ваше мнение о ценности публичной деятельности?

Я привык относиться к проектам так, чтобы стабильно развиваться, продолжая над чем-то работать. В аспирантуре моя идея была бесполезной, но потом она превратилась в код, черновик, чистовик и, наконец, в опубликованную статью. Я считал, что со временем моя работа становилась все более ценной.

С тех пор я понял, что заблуждался. Все, что есть у вас в компьютере, на какой бы стадии готовности оно ни было, ничего не стоит. Если ваша работа не увидела свет, силы потрачены впустую, ведь гораздо более ценно то, что доступно многим. Я понял это благодаря нескольким статьям, которые подготовил в аспирантуре, но так и не опубликовал. Я вложил в них много труда, но мне все время казалось, что они еще не совсем готовы. Спустя годы я забыл, что в них было: я не могу их найти и они не принесли миру никакой пользы. Если бы в процессе работы я написал пару сообщений в блоге, отправил пару твитов и, возможно, сделал действительно простой пакет с открытым исходным кодом, все это добавило бы ценности моим усилиям.

Как вы придумываете идеи для своих постов о Data Science?

Я выработал привычку: каждый раз, когда вижу датасет, скачиваю его и быстро просматриваю, запускаю пару строк кода, чтобы прочувствовать данные. Так можно лучше понимать Data Science: поработав над достаточным количеством проектов, вы начнете понимать, в каких фрагментах данных может быть скрыто что-то интересное, а какие, возможно, даже не стоит обрабатывать.

Мой вам совет: когда у вас появится возможность проанализировать данные — даже если они не относятся к вашей текущей работе или вы считаете, что вас это не заинтересует, — потратьте буквально несколько минут на их просмотр и оцените, что там может быть интересного. Выберите датасет, решите, сколько времени готовы потратить, проанализируйте все, что сможете, а затем опубликуйте результаты. Ваш пост может быть неидеальным, вы можете найти не все, на что рассчитывали, и ответить не на все вопросы, на которые хотели, но, задавшись целью писать статью о каждом наборе данных, вы сможете вырабатывать эту привычку.

Ваш последний совет новичкам и джуниорам в Data Science.

Не беспокойтесь о том, что нужно быть в курсе новейших технологий. Когда вы начинаете работать в сфере Data Science и МО, возникает соблазн начать с глубокого обучения или других продвинутых методов. Важно помнить, что они были разработаны для решения некоторых наиболее сложных задач в этой области и сейчас вам вовсе не обязательно применять их для решения тех задач, с которыми

вы столкнетесь как начинающий специалист. Начните с простого преобразования и визуализации данных, программирования с применением широкого набора библиотек, используйте статистические методы, такие как проверка гипотез, классификация и регрессия. Для начала следует освоить эти методы и только потом переходить к передовым технологиям.

Итоги

- Наличие портфолио в репозитории GitHub и блоге может помочь вам получить работу.
- Есть много источников, где можно найти хорошие датасеты для стороннего проекта; главное — выбрать что-то интересное и нестандартное.
- Не обязательно посвящать блог только сторонним проектам; можно также поделиться учебными пособиями или рассказать о своем опыте, полученном в буткемпе или во время конференции.

Материалы к главам 1–4

Книги

Practical Data Science with R, 2nd ed., Nina Zumel and John Mount (Manning)

Эта книга представляет собой введение в Data Science, где в качестве основного инструмента используется язык R. Она станет отличным дополнением к нашей книге, потому что в ней гораздо глубже рассматриваются технические аспекты работы. Она объясняет, как выбирать датасеты, придумывать к ним правильные вопросы и интерпретировать результаты.

Doing Data Science: Straight Talk from the Frontline, Cathy O'Neil and Rachel Schutt (O'Reilly)

Это еще одно введение в Data Science со смесью теории и практики. Это не набор кейсов: книга широко рассматривает сферу Data Science и пытается подойти к ней с разных сторон.

R for Everyone, 2nd ed., Jared Lander, и *Pandas for Everyone*, Daniel Chen (Addison-Wesley Data and Analytics)

Это две книги из серии Addison-Wesley Data and Analytics. Они охватывают использование R и Python (через pandas) от базовых функций до расширенного анализа и решения DS-задач. Эти книги станут отличным ресурсом для тех, кто нуждается в помощи по этим темам.

Think Like a Data Scientist: Tackle the Data Science Process Step-by-Step, Brian Godset (Manning)

Это книга по Data Science для новичков. В ней рассказывается о том, как на самом деле устроена работа в этой сфере. Последовательно изложены особенности постановки задач, создания плана, процессов решения и демонстрации результатов. Эта книга лучше всего подойдет тем, кто разбирается в технических аспектах Data Science, но плохо знаком с работой над долгосрочным проектом.

Getting What You Came For: The Smart Student's Guide to Earning an M.A. or a Ph.D., Robert L. Peters (Farrar, Straus and Giroux)

Если вы решили получить степень магистра или кандидата наук, вас ждет долгий и изнурительный путь. Никто прямо не учит тому, как нужно сдавать экзамены, получать квалификацию, проводить исследования или быстро решать сложные задачи. И хотя эта книга довольно старая, ее уроки по успешному обучению остаются актуальными.

Bird by Bird: Some Instructions on Writing and Life, Anne Lamott (Anchor)

Bird by Bird («Птичка за птичкой») — не только руководство по усовершенствованию навыков письма, но и отличный путеводитель по жизни. Название взято из истории о птичках. Однажды брат автора бегал в панике от того, что должен был написать отчет о птицах за три месяца, но оставил все на последний вечер перед сдачей. Тогда их отец сказал ему: «Птичка за птичкой, приятель. Просто сделай все по порядку». Если вы боретесь с перфекционизмом или пытаетесь понять, о чем можно написать, возможно, эта книга — именно то, что вам нужно.

Блоги

«Bootcamp rankings», Switchup.org

https://www.switchup.org/rankings/best-data-science-bootcamps

Switchup дает список 20 лучших буткемпов, составленный на основе отзывов студентов. Хотя вы можете не доверять обзорам и отзывам, с этого блога все равно можно начать выбор подходящего буткемпа.

«What's the Difference between Data Science, Machine Learning, and Artificial Intelligence», David Robinson

http://varianceexplained.org/r/ds-ml-ai

Если вы не понимаете, в чем разница между Data Science, машинным обучением и искусственным интеллектом, то этот пост расскажет вам о хорошем способе отличать эти направления. При том что общепринятых определений не суще-

ствует, нам нравится система, предложенная автором: Data Science дает инсайты, МО — прогнозы, а ИИ — действия.

«What You Need to Know before Considering a PhD», Rachel Thomas

https://www.fast.ai/2018/08/27/grad-school

Если вы считаете, что без кандидатской степени вам не стать дата-сайентистом, сначала прочтите этот блог. Автор объясняет, как дорого вам может обойтись этот шаг (с точки зрения как потенциальных психологических проблем, так и альтернативных карьерных издержек), и развенчивает миф о необходимости степени для проведения передовых исследований в области глубокого обучения.

«Thinking of Blogging about Data Science? Here Are Some Tips and Possible Benefits», Derrick Mwiti

http://mng.bz/gVEx

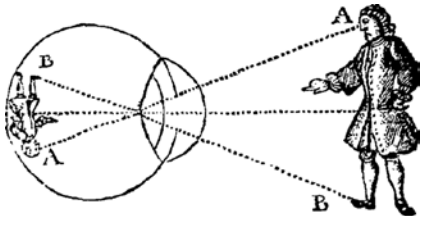
Если глава 4 не убедила вас в преимуществах ведения блога, возможно, вас убедит эта статья. Автор также предлагает несколько отличных советов по созданию более интересных публикаций, в том числе с помощью маркированных списков и новых датасетов.

«How to Build a Data Science Portfolio», Michael Galarnyk

http://mng.bz/eDWP

Это отличная подробная статья о создании портфолио. Автор рассказывает не только о типах проектов, которые стоит в него включать (или нет), но и о способах добавить их в резюме и поделиться с другими.

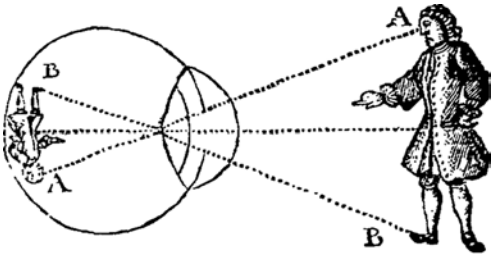
Часть 2 Как попасть в Data Science



Теперь, когда у вас есть весь арсенал для работы в Data Science, пришло время заняться ее поиском. Эта часть книги охватывает все, что вам нужно знать для успешного устройства в компанию, начиная с рассмотрения открытых вакансий и заканчивая переговорами и принятием оффера. Процесс поиска работы в Data Science имеет некоторые уникальные особенности из-за специфики этой сферы. На примере кейса мы научим вас фильтровать сотни вакансий, где требуются *дата-сайентисты*, и что конкретно нужно компаниям. И хотя эта часть книги особенно полезна для тех, кто еще не работал в сфере, материал может также освежить память младшим или старшим специалистам по данным.

Глава 5 охватывает поиск работы в Data Science и то, как не потеряться среди огромной массы вакансий. Из главы 6 вы узнаете, как составлять убедительное резюме и сопроводительное письмо. В главе 7 мы поговорим о том, чего следует ожидать от интервью и как к нему подготовиться. Мы рассмотрим все этапы — от первого телефонного звонка до финальной встречи. Из главы 8 вы узнаете, что говорить при получении оффера от компании: следует ли его принимать, а также зачем и как вести переговоры.

5



Поиск: как определиться с подходящей работой

В этой главе

- Поиск подходящих открытых вакансий.
- Расшифровка описания вакансий для понимания истинной сути должности.
- Выбор вакансий для отклика.

У вас есть навыки и портфолио; единственное, чего вам не хватает, — это работы! Однако вы должны быть готовы к тому, что процесс поиска займет некоторое время. Даже если все сложится удачно, вам понадобиться не менее месяца, начиная от подачи заявки до получения оффера, а чаще — несколько месяцев. Но с помощью лучших практик, перечисленных в этой главе, мы надеемся сделать для вас этот процесс максимально безболезненным.

Здесь мы подробно расскажем о том, как искать работу в Data Science. Во-первых, мы рассмотрим все возможные для этого места, чтобы вы ненароком не сузили свой выбор. Затем мы покажем, как расшифровать описания требований к кандидатам, чтобы вы понимали, какие навыки вам действительно нужны (спойлер: не все) и каким должностям они могут соответствовать. Наконец, мы научим, как выбрать наиболее подходящее место с учетом ваших знаний и опыта, а также типов компаний, о которых мы говорили ранее.

5.1. Поиск работы

Прежде чем думать о составлении «идеального» резюме и сопроводительного письма, вам нужно знать, куда их отправить! Начните с сайтов вакансий, таких как LinkedIn, Indeed и Glassdoor. Стоит выбрать несколько веб-сайтов, потому что не все компании публикуют свои вакансии на каждом. Если вы относитесь к группе недостаточно представленного меньшинства в сфере технологий, поищите специальные сайты, такие как POCIT и Tech Ladies, которые предназначены для цветных людей и женщин в этой области соответственно. Место поиска зависит от вида работы, на которую вы претендуете, например бывают сайты объявлений для определенных типов компаний, таких как AngelList для стартапов или Dice для технологических организаций.

Обязательно сделайте поиск расширенным. Как мы уже говорили, вакансии в Data Science имеют множество различных названий. Одна и та же должность называется по-разному в зависимости от организации, а некоторые компании даже называют специалистов по-новому, оставляя прежний круг обязанностей. Например, раньше это были аналитики данных, а через год их переименовали в дата-сайентистов! Вот некоторые примеры, с которыми вы можете столкнуться (рис. 5.1):

- *Аналитик данных*. Эта должность часто является первой ступенью и может стать отличным стартом для начала работы в этой сфере, если у вас нет степени в области естественных наук, технологий, инженерии или математики (STEM) и опыта. Как мы позже расскажем в разделе 5.1.1, следует очень осторожно подходить к этой должности и убедиться, что в обязанности будут входить программирование и статистика или МО.
- *Специалист по количественному анализу, аналитик продукции, аналитик-исследователь и прочие виды аналитиков*. С точки зрения обязанностей эти должности еще разнообразнее, чем аналитик данных. Вы можете выполнять ту же работу, что и дата-сайентисты в других компаниях, равно как и коротать дни, работая с устаревшими таблицами Microsoft Excel.
- *Инженер по машинному обучению*. Как следует из названия, эти вакансии ориентированы на МО в Data Science и обычно требуют опыта работы инженером. Если у вас есть степень в области компьютерных наук или вы работали инженером-программистом, такая должность может отлично вам подойти.
- *Инженер-исследователь*. Для этих должностей часто требуется кандидатская степень, хотя это может быть обсуждаемо, если у вас есть степень магистра в области компьютерных наук, статистики или смежной области.

Попробуйте сначала просто поискать *данные* на одной из платформ по трудоустройству и потратить час на чтение объявлений о вакансиях: так вы поймете, какие отрасли есть в вашем регионе и какие типы вакансий открыты. Кроме того,

Рис. 5.1. Должности, в названии которых не упоминается работа с данными, но которые вы можете встретить при поиске

у вас появится тактика для быстрой фильтрации новых объявлений. Просмотр наиболее подходящих вам должностей вместо всех доступных вакансий сузит область поиска до приемлемых масштабов. Не обращайте внимания на название вакансии: лучше используйте описание, чтобы оценить, насколько она вам подходит.

Будьте предельно осторожны, относясь к поиску работы как к игре в числа. В списке будут сотни вакансий, если вы ищете их в большом технологически развитом городе, таком как Нью-Йорк или Сан-Франциско, или в нескольких городах сразу. Поиск может быстро стать навязчивой идеей, ведь это простой способ почувствовать себя продуктивным («Сегодня я просмотрел 70 вакансий!»). И так же, как в случае с Twitter и Facebook, постоянная проверка обновлений может вызывать привыкание. Заниматься этим чаще, чем каждые три-пять дней, как правило, бесполезно. Если же делать это раз в месяц, то можно упустить хорошую возможность, но все же ни одна компания не закроет вакансию (которая фактически открыта) через два дня после ее размещения на сайте.

Если вас интересует работа в конкретных компаниях, ищите вакансии на их сайтах. Так как вы будете ориентироваться на разные названия должностей, проверяйте разные подразделения. В некоторых компаниях вакансии в Data Science можно найти в финансовом, инженерном и других отделах, и если вы не заглянете туда, то не найдете их.

> **ДЖУНЫ** Когда будете искать работу, ищите должности с названиями «начинающий специалист», «Junior», «младший сотрудник» и «начальный уровень». Также обратитесь за помощью в центр профориентации и посещайте разные ярмарки вакансий в вашем учебном заведении.

5.1.1. Расшифровка описания вакансий

Когда вы начинаете читать описания вакансий, может показаться, что вся работа в Data Science попадает в одну из двух категорий:

- *Бизнес-аналитик*. На этой должности вы будете использовать инструменты бизнес-аналитики, такие как Excel и Tableau, и, возможно, немного SQL, но, как правило, вы не будете писать код. Если вы хотите улучшить навыки программирования, знания в области статистики и инженерии данных или расширить набор инструментов для машинного обучения, эти вакансии вам не подходят.
- *Единорог*. С другой стороны, есть целый ряд вакансий с такими требованиями: наличие кандидатской степени в области компьютерных наук, опыт работы дата-сайентистом более пяти лет; эксперт в области передовой статистики, глубокого обучения и общения с деловыми партнерами; имеет опыт выполнения широкого круга обязанностей, от машинного обучения на производственном уровне до создания информационных панелей и проведения A/B-тестов. Такие описания обычно означают, что компания сама не знает, кого ищет, и ожидает, что специалист по данным решит все ее проблемы самостоятельно.

Но не спешите расстраиваться: даем слово, что бывает и другая работа. Лучше думать о вакансии с точки зрения квалификации. Компании нужен кто-то для создания нового отдела, при том что никакой инфраструктуры конвейера данных пока что нет? Или же она ищет пятого сотрудника в существующую продуктивную DS-команду с надеждой, что новый человек сможет сразу же внести свой вклад в работу, но при этом от него не ждут, что он будет одновременно отлично управлять данными, иметь навыки делового общения и к тому же разрабатывать ПО? Чтобы разобраться, внимательно прочтите описание должности и постарайтесь понять, что на самом деле ищет работодатель. Представьте себе сайт, на котором пристраивают животных в добрые руки. Вы увидели фото кошки Дыньки и описание: «Любит интересоваться, как у вас дела». Вы должны понимать: на самом деле это означает, что кошка постоянно мяукает и требует внимания — сомнительная перспектива.

В описаниях вакансий есть клише, на которые также стоит обратить внимание. Фраза: «Хорошо поработал — хорошо отдохнул» означает, что вам придется много работать, но зато вы сможете посещать неформальные корпоративные мероприятия (например, поход в бар). Или, например, если ищут «инициативного и независимого» сотрудника, значит, вам будут мало помогать. Умение читать между строк поможет найти подходящую работу.

Прежде всего, следует иметь в виду, что список требований — это, скорее, пожелания, так что здесь есть место для маневра. Если вы отвечаете 60 % требований (возможно, вам не хватает года до необходимого опыта работы или вы не работали с одним из компонентов технологического стека компании), но в остальном подходите, смело откликайтесь на вакансию. Не придавайте существенного значения таким формулировкам, как «будет вашим плюсом». Кроме того, требования к многолетнему опыту работы лишь указывают на необходимые навыки;

если вы программировали во время учебы в магистратуре, этот опыт может быть учтен. Однако если вы начинающий дата-сайентист и пришли из сферы маркетинга, то идея откликаться на должность старшего специалиста, где от кандидата требуется пятилетний опыт работы, отличные знания Spark и Hadoop и умение развертывать модели МО, — не лучший вариант, если, конечно, вы не хотите потратить время впустую. В этом случае компания ищет кандидата с другим уровнем квалификации.

> **ТРЕБОВАНИЯ К ОБРАЗОВАНИЮ** Многие вакансии в Data Science требуют наличие образования по «количественной дисциплине» (например, статистика, инженерия, информатика или экономика). Можно ли откликнуться, если соответствующего диплома у вас нет? Как правило, да. В главе 6 мы обсудим это подробнее, но если вы посещали занятия по этим направлениям (включая буткемпы или онлайн-обучение), то можете это указать. Если вы последовали совету из главы 4 — создали портфолио и публиковали статьи в блоге, можете показать эти проекты работодателям как доказательство, что справитесь с работой.

Один из нюансов вакансий в этой сфере заключается в том, что разные слова могут означать одно и то же. В машинном обучении и статистике это встречается довольно часто. В одной компании требуется опыт регрессионного анализа или классификации, в другой — опыт контролируемого обучения, но в целом эти термины эквивалентны. То же самое касается A/B-тестирования, онлайн-экспериментов и рандомизированных контрольных испытаний. Если вы не знаете термин, загуглите его; вы можете обнаружить, что уже делали такое, просто оно по-другому называлось! Если вы не работали с какой-то технологией, заявленной в вакансии, вспомните, сталкивались ли вы в принципе с чем-то подобным. Если в списке упоминаются, например, Amazon Web Services (AWS), а вы работали с Microsoft Azure или Google Cloud, значит, у вас есть навыки работы с сервисами облачных вычислений.

У способности читать между строк есть еще один плюс: умение разглядеть тревожные сигналы (см. раздел 5.1.2). Ни одна компания не назовет предлагаемую работу плохой. Чем раньше вы распознаете, что с вакансией что-то не так, тем лучше, поэтому стоит отмечать любые странности в описании должности.

5.1.2. Поиск тревожных сигналов

Поиск работы — это игра на двоих. В процессе поиска вам может казаться, что вся власть сосредоточена в руках компаний, которым нужно доказать, что вы достойны работы у них. Но вы (да, вы!) тоже можете быть избирательными. Попасть в токсичные условия или на ужасно скучную работу — не лучшая перспектива.

И пусть вы не всегда сможете определить это по описанию, можно обратить внимание на несколько признаков:

- *Отсутствие описания*. Первый звоночек — это наличие только списка требований и отсутствие описания компании или обязанностей. Такие организации забывают, что прием на работу — это процесс, в котором участвует две стороны, и им попросту нет до вас дела. Либо компании, возможно, купились на ажиотаж в Data Science и просто хотят иметь собственных специалистов, ничего не вкладывая в создание условий для их эффективной работы.
- *Слишком широкие требования*. Второй звоночек — это вышеупомянутое описание единорога (см. раздел 5.1.1). Несмотря на то что это пример крайности, вам следует с осторожностью относиться к любому описанию обязанностей, в котором две или три сферы деятельности (поддержка принятия решений, аналитика и машинное обучение) указаны в качестве основных. Можно ожидать, что у кандидата есть базовые знания по каждой из них, но ни один человек не сможет выполнять все эти функции как эксперт. Даже если бы кто-то и мог, ему попросту не хватило бы времени на все.
- *Несоответствия*. Также обратите внимание на несоответствие между требованиями к кандидату и описанием обязанностей. Работодатель требует опыта в области глубокого обучения, но в обязанностях фигурирует создание информационных панелей, общение с заинтересованными сторонами и работа с экспериментальными программами? Если это так, компании нужен кто-то, кто умеет пользоваться наиболее трендовыми инструментами или она просто ищет «престижного» дата-сайентиста с докторской степенью Стэнфорда в области искусственного интеллекта, хотя на самом деле эти специализированные знания не пригодятся.

5.1.3. Большие надежды

Хотя вы должны понимать, к чему стремиться, не следует гнаться за совершенством. Начинающие дата-сайентисты иногда видят свой путь развития примерно так: «Шаги 1–98: изучить Python, R, глубокое обучение, байесовскую статистику, облачные вычисления, A/B-тестирование, D3. Шаг 99: получить работу. Шаг 100: профит». Конечно, это преувеличение, но весь ажиотаж вокруг Data Science — это идеализация реальной ситуации. В конце концов, специалист по данным — «лучшая работа в Америке» (http://mng.bz/pyA2) с шестизначной зарплатой и высоким уровнем удовлетворенности. Вы можете вообразить, будто каждый день придется решать интереснейшие задачи с умнейшими коллегами. Необходимые данные всегда будут доступны и очищены, а любые проблемы будут моментально решены командой инженеров. Работа будет в точности соответствовать описанию в вакансии, и вам никогда не придется заниматься скучными вещами.

К сожалению, такой сценарий — утопия. Надеемся, что часть I этой книги убедила вас в том, что для устройства на работу не обязательно знать все; так же и компании не обязаны быть идеальными единорогами. По определенным причинам эта книга не заканчивается на том, как вас берут на должность. И хотя стать дата-сайентистом — это уже большое достижение, которым вы должны гордиться, учтите, что вы попали в ту сферу, в которой придется постоянно учиться. Вы столкнетесь с нерабочими моделями, с политикой компании, которая сводит на нет всю проделанную за последний месяц работу, а еще вам придется потратить не одну неделю на переговоры с инженерами и продакт-менеджерами, чтобы собрать необходимые данные.

Особенно легко идеализировать хорошо известные компании. Может быть, вы пришли на собеседование и кто-то из сотрудников поразил вас. Возможно, вы уже несколько месяцев читаете блог этого человека и знаете, насколько он продвинутый в своей области. А может, вы прочитали в какой-то статье, что в офисе компании можно спать, есть блюда высокой кухни и обниматься с милыми собачками. То, что привлекло вас, наверняка заинтересовало и других начинающих специалистов; большинство этих компаний получают сотни откликов и могут установить более высокую планку, чем в действительности необходимо для выполнения задач. В любом случае речь может идти о работе в другом подразделении, а сама должность окажется невероятно скучной.

Даже с вполне реалистичными ожиданиями вы вряд ли сразу получите работу мечты. Легче перейти на эту должность внутри компании или включить анализ данных в свои текущие обязанности; даже если вы в конечном итоге собираетесь уйти из своей сферы, для начала все же потребуется взять на себя другую роль, чтобы прокачать навыки. Это не значит, что у вас не должно быть определенных требований и предпочтений; просто проявите гибкость. Менять работу в сфере технологий даже через год или два — абсолютно нормально, так что вы не подписываетесь на это на ближайшие 15 лет. Но вы не можете точно знать заранее, чего хотите, и даже на плохой работе можно научиться чему-то полезному, так что сильно не переживайте.

5.1.4. Посещение митапов

Хотя порталы по трудоустройству — это распространенный способ найти открытые вакансии, как правило, давать отклик через них не особо эфективно. В главе 6 мы расскажем, что холодные онлайн-заявки часто дают очень низкий процент обратной связи. Согласно опросу, проведенному Kaggle в 2017 году (https://www.kaggle.com/surveys/2017), уже трудоустроенные дата-сайентисты ищут и получают работу через рекрутеров, друзей, членов семьи и коллег. Отличный способ собрать такую сеть контактов — посещать митапы.

Митапы обычно представляют собой личные встречи, которые проводятся по вечерам в будни. Как правило, на них выступает один докладчик или целая группа специалистов, представляющих заявленную тему мероприятия. Митапы должны быть бесплатными или иметь символическую стоимость, которая иногда уходит на еду. На некоторых может быть только двадцать человек; другие могут вмещать до трехсот. Одни участники проводят митапы каждый месяц, другие встречаются всего несколько раз в год. Где-то они остаются для дальнейшего общения или идут в ближайший бар; а в других местах все держится исключительно в рамках лекций. Одни митапы посвящены очень специфическим направлениям, таким как расширенная обработка естественного языка на Python, другие в этом месяце могут обсуждать введение во временные ряды, а в следующем — продвинутые модели глубокого обучения. Чтобы понять, что вам больше нравится, лучше посетить несколько форматов. Тема, безусловно, важна, но вы должны найти то место, где почувствуете себя желанным гостем и получите удовольствие от общения с другими участниками. Почти у всех митапов есть учетная запись на https://www.meetup.com, так что можно выполнять поиск по теме. Наберите в строке поиска «Data Science», «машинное обучение», «Python, R» или «аналитика», чтобы найти подходящие мероприятия в вашем регионе.

Вначале многих митапов люди заявляют о наборе сотрудников. Подойдите к ним и пообщайтесь: подбор персонала — это часть их работы, и даже если текущие вакансии вам не подходят, вам могут дать хороший совет или предложить альтернативные места для поиска.

Вы также можете встретить другого участника митапа, который работает в интересующей вас компании или отрасли. Спросите, есть ли у него время ответить на несколько вопросов, чтобы вы могли больше узнать об этой области. Такой разговор не является (или, скорее, не должен быть) пассивно-агрессивным способом поиска работы, скорее это отличная возможность получить взгляд изнутри и совет от специалиста. В главе 6 мы поговорим о преимуществах получения рекомендаций, однако не советуем просить о них людей, с которыми едва познакомились. Вы хотите слишком многого от человека, который вас толком не знает; кроме того, никому не нравится чувствовать, будто его используют. Если кто-то расскажет вам о вакансии в своей компании и предложит вас порекомендовать — отлично! Но даже если этого и не произойдет, вы все равно сможете многое извлечь из такого разговора.

Посещение митапов полезно и по другим причинам. Во-первых, они позволяют находить единомышленников рядом с вами. Если вы переехали в другой город или только что окончили колледж, то можете почувствовать себя чужим. Митапы — отличный шанс сделать карьеру и выстроить круг общения. Можно воспользоваться преимуществами митапов, чтобы просто пообщаться или завязать знакомства, которые могут помочь вам с чем угодно, начиная от конкретных

вопросов по анализу данных до получения рекомендаций при поиске работы или менторинга. Учтите, что некоторые митапы публикуют записи выступлений в интернете, а другие — нет, поэтому личное присутствие — это единственный способ поучаствовать и увидеть все своими глазами.

К сожалению, у митапов есть несколько недостатков. Может быть страшно участвовать в мероприятии, где у всех есть опыт и/или все друг друга знают. Синдром самозванца никто не отменял, но вы обязаны его побороть, так как мест с дружественной атмосферой все же больше, чем просто хороших митапов. Наконец, хотя такие мероприятия помогают увидеть кухню Data Science изнутри, они могут быть закрытыми или неразнообразными в плане состава участников: это зависит от приветливости организаторов и от связи митапа с разными сообществами.

5.1.5. Использование социальных сетей

Если вы не живете в городе или рядом с ним, то, возможно, не сможете найти DS-митапы поблизости. В этом случае вам отлично подойдут Twitter и LinkedIn. Если вы подписаны на нескольких известных дата-сайентистов, то по их хештегам или ретвитам сможете найти других людей, на которых можно подписаться. Вы также можете начать создавать личный бренд.

Нам нравится использовать Twitter несколькими способами:

- *Делиться своей работой*. После написания отличной статьи в блоге вы хотите, чтобы другие люди узнали о ней! Самореклама со ссылками на свою работу с кратким описанием — совершенно нормальная практика.
- *Делиться работой других людей*. Вы прочитали что-нибудь классное? Пакет сохранил вам кучу нервов? Увидели в выступлении слайд, который оказался особенно полезным? Если да, помогите другим людям. В главе 6 мы обсуждаем один из лучших способов связаться с кем-либо: рассказать о пользе, которую принесла вам работа этого человека. Отметить автора в своем посте с помощью хештега — отличный способ привлечь его внимание. Если вы делитесь выступлением, проверьте, есть ли у конференции или митапа свой хештег; с ним ваш твит станет заметнее.
- *Обращаться за помощью*. У вас есть задача, на которой вы застряли и не смогли решить даже с помощью Гугла? Вполне вероятно, что кто-то другой столкнулся с той же проблемой. В зависимости от типа задачи можно пойти на определенные форумы или веб-сайты, задать вопросы или кинуть клич, используя соответствующий хештег.
- *Обмениваться советами*. Не все нужно непременно публиковать в блоге, но если у вас есть совет по быстрому решению задачи — поделитесь им. Вы можете считать, что вопрос того не стоит и все и так все знают, но помните:

начинающие специалисты многого не понимают. Даже люди, которые годами использовали определенный язык, могут не знать какой-то новый способ решения.

В соцсетях можно также писать, что вы ищете работу, и спрашивать других о потенциальных вариантах. Даже если у вас еще не сформировалась обширная сеть контактов, ваши друзья, бывшие одноклассники и коллеги могут знать о вакантных должностях в компаниях. Этот подход обычно эффективнее работает на платформах, объединяющих дата-сайентистов, например на LinkedIn или Twitter, но даже на Facebook можно найти полезные контакты.

В начале карьеры часто кажется, что связи есть только у тех, кто уже работает с данными. Секрет в том, что их нужно искать задолго до поиска работы. Чем больше вы сможете выйти из своей зоны комфорта и общаться с людьми на конференциях, митапах, в академических учреждениях, на барбекю и так далее, тем более подготовленными вы будете при следующем поиске вакансий.

Запасной аэродром

Распространенная ошибка при поиске работы — возложение больших надежд только на одну вакансию и отказ от собеседований в других компаниях. А вдруг вам откажут? Вы же не хотите снова начинать поиск с нуля. На каждом этапе у вас должно быть несколько запасных вариантов: отправленные отклики, отсеивание рекрутером, тестовое задание и собеседование. Не считайте, что все закончилось, пока не примете оффер в письменной форме.

Наличие нескольких вариантов также помогает проще реагировать на отказ. Когда вы ищете работу, отказы почти неизбежны, а не принимать их на свой счет или не воспринимать их как показатель вашей ценности иногда непросто. В некоторых случаях вам могут даже не сообщить об отказе — вам просто не ответят. Но многие причины отказа зависят вовсе не от вас. Компания могла закрыть вакансию, никого не наняв, взять кандидата внутри организации или принять того, кто откликнулся на порядок раньше вас. Получить отказ всегда обидно, особенно от компании, которая вас действительно заинтересовала, поэтому потребуется немного времени, чтобы осмыслить свои чувства. Наличие же других вариантов поможет вам сохранить мотивацию и двигаться дальше.

Наконец, имея несколько возможных вариантов, вам проще отказать работодателю. Вы могли пройти собеседование с рекрутером и справиться с тестовым заданием только для того, чтобы в итоге обнаружить, что в компании нет инженеров по обработке данных, в команде по анализу данных всего несколько человек (и это при том, что компания крупная), или же понять, что реальные требования сильно отличаются от заявленных в описании вакансии. Несмотря на то что ждать идеальных условий не стоит (поскольку их попросту не существует), у вас наверняка есть несколько жестких требований, придерживаться которых гораздо легче, если вы понимаете, что в других организациях они будут соблюдены.

5.2. На какие вакансии откликаться

Сейчас у вас уже должен быть список как минимум из десятка вакансий, которые вам в той или иной мере интересны и могут подойти. Вы откликнетесь на все сразу?

Некоторые действительно откликаются на десятки или даже сотни вакансий. Они пытаются постелить себе соломки, полагая, что если шанс получить отклик на любую вакансию равен 10 %, то направление откликов в как можно большее количество компаний даст больше обратной связи. Но это не так: при ограниченных ресурсах и времени и распылении на 100 предложений вместо 10 шансы на положительный исход только падают. В главе 6 мы поговорим о том, как адаптировать отклики под каждую вакансию, но учтите, что метод работает только если вы будете избирательны. Если вы решите откликнуться на объявления 50 компаний, вы практически обречены.

> **R И PYTHON** Следует ли откликаться на вакансию, если в требованиях указано знание Python, а вы знаете R, или наоборот? Зная один язык, вам, безусловно, будет проще освоить другой, однако вам уже и так предстоит многому учиться, например работать со стейкхолдерами, осваивать внутреннюю политику, статистику, датасеты и так далее. Даже если бы вас взяли, изучение нового языка в дополнение ко всему остальному может сложно даваться. Поэтому мы обычно рекомендуем откликаться только на те вакансии, где указан основной язык, который вы знаете. Держите руку на пульсе: если в описании сказано, что знание одного из языков является плюсом, а другой в требованиях не указан, это может означать, что на самом деле программировать вы не будете. Наконец, некоторые вакансии требуют знания обоих языков. Не расслабляйтесь: обычно это требование означает, что специалисты будут использовать любой язык, который знают, а это может усложнить совместную работу. Такая вакансия может вам подойти, но не забудьте уточнить во время собеседования, по какому принципу выбирается язык. Если в команде из 20 человек вы будете одним из двух, использующих Python, улучшить навыки программирования будет сложно.

Вспомните, что вы узнали из первых двух глав о типах DS-компаний и видах работы. Хотите ли вы заниматься всеми направлениями Data Science, настраивая в этом месяце систему рекомендаций, а в следующем создавая модель ценности? Если да, вероятно, вам следует попробовать себя в фирме, которая начала заниматься этим недавно, поскольку в более устоявшихся организациях будут специализированные должности. С другой стороны, в крупных технологических компаниях есть легионы дата-инженеров, поэтому всегда можно легко и быстро получить стандартные данные.

Что-то будет очевидно на основании фактов о компании; например, у стартапа из десяти человек явно не будет развитой системы обработки и анализа данных. Но как узнать больше?

Во-первых, посмотрите, есть ли у организации блог о Data Science. Обычно он есть только у технологических компаний, но изучение таких публикаций помогает понять, чем на самом деле занимаются специалисты по данным. Включите свои положительные отзывы о блоге в сопроводительное письмо (см. главу 6) — это обязательно оценят. Если вы прежде никогда не слышали о фирме, изучите ее веб-сайт. Когда вы знаете, чем занимается компания и как она зарабатывает деньги, вы можете предугадать, каким типом работы с данными придется заниматься. Наконец, если организация вам действительно интересна, посмотрите, есть ли у кого-нибудь из ее сотрудников блог, в котором они рассказывают о ней или о своей работе.

Читая о компании, не забывайте о том, что важно конкретно для вас. Хотите иметь возможность работать удаленно? А что насчет длительности отпуска? Выделяет ли компания время и деньги для поездок на конференции? Кроме того, информация, которую фирма дает о себе, может рассказать вам о ее ценностях. Настольный футбол, пиво в офисе и доставка обедов? Там, вероятно, будет много молодых сотрудников. Упор на гибкий график работы или отпуск по семейным обстоятельствам? Эта компания, скорее всего, будет хорошо относиться к родителям. В главе 8 мы расскажем, как договариваться не только о зарплате, но на этом этапе вы как минимум можете понять, соответствуют ли преимущества компании вашим приоритетам.

Теперь, когда у вас есть приемлемый перечень подходящих вакансий, пора откликаться! В главе 6 мы расскажем, как написать отличное резюме и сопроводительное письмо, а также адаптировать их для каждой компании.

5.3. Интервью с Джесси Мостипак, developer advocate в Kaggle

Джесси Мостипак (Jesse Mostipak) работала в области молекулярной биологии и была учителем в государственной школе, прежде чем влюбилась в некоммерческое направление Data Science. На момент этого интервью она занимала должность управляющего директора Data Science в Teaching Trust. Вы можете найти ее статьи о некоммерческом DS, советы по изучению R и другие темы на ее веб-сайте https://www.jessemaegan.com.

С чего лучше начать поиск работы?

Подумайте о том, насколько вам важно название должности. Если вы отбросите предрассудки и сосредоточитесь непосредственно на обязанностях, перед вами откроется гораздо больше перспектив. Некоторые ключевые слова, которые не указаны в названии вакансии «дата-сайентист», — это *анализ*, *аналитика* и *данные*. С большим числом фильтров можно найти заголовок вакансии вроде «исследова-

ние и оценка», где по функционалу вы подходите на должность, которую никогда бы не нашли, будь у вас в поиске только стандартное название.

Сосредоточьтесь на том, чем именно хотите заниматься как специалист по данным. Мне, например, неинтересно рассчитывать рентабельность инвестиций по количеству кликов на веб-сайтах. Я спросила себя: «Чего я хочу? Какие организации соответствуют моим интересам?» Мне было интересно поработать с девочками-скаутами (Girl Scouts): оказалось, что они искали аналитика, и взяли меня. То же самое произошло с Teaching Trust, когда я захотела больше заниматься образованием.

Как создать свою сеть контактов?

Пытаясь попасть в сферу Data Science, я делала много бесполезных вещей. Я репостила в Twitter каждую прочитанную мной статью о DS, размещая 20 бессмысленных постов в день. Решите, с кем вы хотели бы встретиться, почему и что ценного вы приносите в эти взаимоотношения. Подумайте о личном бренде, убедитесь, что ваш образ в интернете и в социальных сетях соответствует действительности. Что касается меня, я поняла, что не могу быть идеальным специалистом и что нет смысла ждать, пока я все выучу, прежде чем вести соцсети, ведь этого никогда не произойдет. Вместо этого я решила рассказывать о том, что я изучаю, и быть открытой. Именно так я и построила свою сеть.

Что делать, если вы не уверены, что стоит откликаться на вакансию?

Если вы развиваете навыки, умеете проводить анализ на Python или R и владеете базовыми знаниями, вам следует сосредоточиться на том, как научиться рисковать и терпеть неудачи. Как дата-сайентист вы должны к этому привыкнуть. Если вы боитесь рискнуть, откликнувшись на вакансию, то что же будет, когда вы рискнете в отношении модели, а она не сработает? Смиритесь с идеей неопределенности и итераций. Вы должны откликнуться и попробовать; если вам откажут, ничего страшного! Мне регулярно отказывают в работе; это часть опыта, и не более того.

Что бы вы сказали человеку, который думает: «Я не отвечаю всем требованиям вакансии?

Как показывает ряд исследований, определенные группы людей считают, что им необходимо иметь 100 %-ную квалификацию, тогда как другие говорят: «Я отвечаю 25 % требований? Отлично, откликаюсь!» Докажите, что вы не хуже, и дерзайте. Но вы также можете неправильно понять описание. Допустим, в списке перечислены отдельные навыки, например 10 лет работы с базами данных SQL. Вы можете подумать: «У меня нет такого опыта; я семь лет работаю с Microsoft Access». Но я бы сказала, что это навыки широкого применения. Вы как кандидат должны сказать себе: «Возможно, у меня нет именно этого навыка, но зато есть

очень похожий на него. Мне нужно посмотреть на SQL, понять, насколько я могу использовать свои навыки, и рассказать этой компании, какие удивительные вещи я делал с Microsoft Access и что они должны меня нанять, потому что я знаю, что могу сделать то же самое с SQL, и не только».

Ваш последний совет начинающим дата-сайентистам?

Вам необходимо развивать коммуникативные навыки и гибкость. Вы должны уметь общаться на всех уровнях своей организации так, чтобы уважать профессионализм людей, с которыми контактируете, а также показывать, что вы здесь для того, чтобы сделать их жизнь проще.

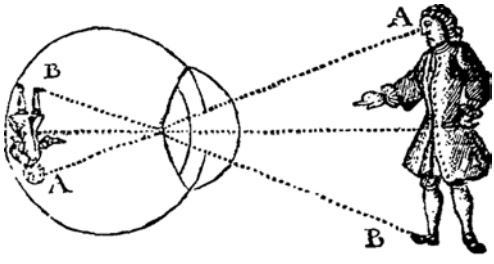
Под *гибкостью* я имею в виду способность сказать что-то вроде: «Я вижу решение этой проблемы или проекта иначе, но я понимаю, что вы имеете в виду. Давайте попробуем вот так, и, возможно, я смогу решить ее вот так». Вы должны быть гибкими, потому что люди еще не до конца разбираются в Data Science. Организации хотят, чтобы у них работали такие специалисты, но когда они их нанимают, то не знают, что с ними делать. Вы радость и отрада организации; если ее потребности изменились, вам необходимо развиваться и адаптироваться, чтобы удовлетворить их наилучшим образом.

Наконец, знайте, что ваши обязанности могут измениться. Вы должны быть в состоянии сказать: «Я не думал, что буду этим заниматься, но как мне к этому приспособиться?» Нельзя говорить: «Я разбираюсь в нейронных сетях лучше всех, но я этим не занимаюсь, поэтому очевидно, что эта работа — дрянь». Вы должны знать, что любая ваша должность будет меняться и развиваться в соответствии с потребностями компании.

Итоги

- Ищите на портале по трудоустройству общие термины, такие как «данные», и уделяйте внимание описанию вакансии, а не ее названию.
- Не волнуйтесь, если понимаете, что не соответствуете требованиям, указанным в описании, на все 100 %.
- Помните, что процесс поиска работы — как улица с двусторонним движением. Обратите внимание на тревожные сигналы и подумайте, каким направлением Data Science хотите заниматься.

6



Отклик на вакансию: резюме и сопроводительное письмо

В этой главе

- Как составить убедительное резюме и написать сопроводительное письмо.
- Как адаптировать отклик под каждую вакансию.

Итак, перед вами список интересных открытых вакансий; пришло время заявить работодателям о своем существовании! Практически каждая вакансия потребует от вас отправки резюме с указанием списка ваших выдающихся навыков и опыта. Для большинства также требуется сопроводительное письмо на одну страницу, в котором вы рассказываете, почему считаете себя подходящим кандидатом. Конечно, легче всего быстро перечислить предыдущие места работы и написать шаблонное письмо, в котором вы выражаете интерес к работе в компании. Однако усилия, которые вы приложите в этом случае, могут стать решающим фактором для приглашения вас на интервью.

В этой главе мы для начала убедимся, что ваше базовое резюме и сопроводительное письмо составлены максимально грамотно, а также поделимся опытом и расскажем о типичных ошибках, которых лучше избегать. Затем мы научим вас «подгонять» основное резюме и сопроводительное письмо под каждую вакансию. Наконец, мы покажем, как сделать так, чтобы ваше тщательно подготовленное резюме попало в руки рекрутера, а не в переполненную стопку с резюме других кандидатов.

В этой главе мы объясним, как важно быстро убедить человека в том, что вы подходите на эту должность. Рекрутеры часто получают сотни резюме на каждую

вакансию в Data Science. Кроме того, поскольку в сфере есть много направлений и должностей, кандидаты, которые откликаются на вакансии, обладают огромным диапазоном навыков. Это еще одно подтверждение тому, что в вашем резюме должно быть сказано что-то вроде: «Эй, хватит просматривать эту огромную кучу с остальными кандидатами, вы уже читаете мое резюме. Вы нашли то, что искали». Но показать, что вы квалифицированный специалист, — задача не из легких.

Хотя нетворкинг и персонализация откликов требуют времени, они дают гораздо лучшие результаты, чем если вы потратите всего час на написание шаблонного сопроводительного письма и резюме, а затем ответите на десятки вакансий одним кликом. У вас будет больше шансов попасть на интервью, так как вы отредактировали свой отклик в соответствии с требованиями компании. А когда вас все же пригласят (тема главы 7), то вы сможете блестяще ответить на распространенный вопрос: «Почему вас заинтересовала эта вакансия?»

6.1. Резюме: основы

Цель вашего резюме не в том, чтобы получить работу, а чтобы попасть на интервью. Рекрутеры, которые нанимают кандидатов, явно не соответствующих требованиям, получают нагоняй от своих руководителей, и наоборот: их хвалят, если подобранные ими сотрудники справляются хорошо. Ваше резюме должно показать, что вы проходите по требованиям, чтобы рекрутер мог спокойно продвигать вас дальше.

Цель резюме заключается вовсе не в перечислении вашего предыдущего опыта, но, к сожалению, многие неопытные соискатели делают именно это. Несмотря на ваше желание избежать пропусков и включить места работы, с которых вы недавно уволились, можете смело удалять опыт, не связанный с Data Science. И даже с солидным опытом за плечами все равно лучше сосредоточиться на выделении наиболее актуальных мест работы. Большинство рекрутеров не станет читать резюме на несколько страниц из-за недостатка времени; а еще они не смогут сказать вашему потенциальному начальнику, на какие его части следует обратить внимание. Никто не скажет вам: «Ну, мы бы вас взяли, но вы не указали, что в старшей школе вы работали спасателем, поэтому мы вам отказываем».

Позже на интервью у вас будет достаточно времени, чтобы подробно рассказать о своих прежних местах работы, образовании и проектах в Data Science. А пока сосредоточьтесь лучше на требованиях, указанных в вакансии. У вас еще будет возможность рассказать о своих выдающихся качествах, которые помогут вам выделиться на фоне других кандидатов, но на первом этапе постарайтесь соответствовать ожиданиям рекрутера.

Исходя из этих соображений, мы рассмотрим основную структуру резюме, расскажем о многочисленных правилах его составления, а также о том, чем его лучше наполнить. Бóльшая часть этой информации применима к любой технической должности в отрасли, но мы максимально сосредоточимся на том, что

свойственно и уникально именно для резюме в области Data Science. Мы также составили резюме, которое вы можете использовать в качестве примера (рис. 6.1).

САРА ДЖОНС

Сан-Франциско, Калифорния 534-241-6264
sarajones@gmail.com linkedin.com/in/sarajones sarajones.github.io github.com/sarajones

ОПЫТ РАБОТЫ

ИЮНЬ 2019 — ПО НАСТОЯЩЕЕ ВРЕМЯ, САН-ФРАНЦИСКО
СТУДЕНТ ПРОГРАММЫ DATA SCIENCE, ПОТРЯСАЮЩИЙ БУТКЕМП

- создала веб-приложение на Python, которое рекомендует вам лучший район Нью-Йорка для проживания в зависимости от бюджета, образа жизни и работы;
- проанализировала 2200 статей на тему бизнеса из газеты New York Times, используя для этого обработку естественного языка, и визуализировала изменение тем с течением времени.

АВГУСТ 2017 — ИЮНЬ 2019, САН-ФРАНЦИСКО
КОНСУЛЬТАНТ ПО ИНВЕСТИЦИЯМ, BIGCO

- создала модель прогнозирования на Python, благодаря которой квартальный доход компании увеличился на 10 %;
- автоматизировала составление еженедельного отчета о рыночных и отраслевых тенденциях.

СЕНТЯБРЬ 2016 — ИЮНЬ 2017, НОВЫЙ ОРЛЕАН
АССИСТЕНТ КАФЕДРЫ ВВЕДЕНИЯ В СТАТИСТИКУ, КРУТОЙ УНИВЕРСИТЕТ

- проводила еженедельные консультации для 60 студентов, средний балл которых составил 4,86 из 5;
- создала и предоставила открытый доступ к учебно-методическому пособию, которое было скачано 1500 раз.

ИЮНЬ 2016 — АВГУСТ 2016, НОВЫЙ ОРЛЕАН
АССИСТЕНТ КАФЕДРЫ ЭКОНОМИКИ, КРУТОЙ УНИВЕРСИТЕТ

- лично провела эксперимент по принятию решений, в котором участвовали 200 человек, а для анализа результатов использовался кластерный анализ на Python;
- результаты исследования были опубликованы в журнале «Экономика».

ОБРАЗОВАНИЕ

ИЮНЬ 2017, НОВЫЙ ОРЛЕАН
БАКАЛАВР ЭКОНОМИКИ, ВТОРОСТЕПЕННАЯ СПЕЦИАЛЬНОСТЬ — СТАТИСТИКА, КРУТОЙ УНИВЕРСИТЕТ
Средний балл успеваемости 3,65/4.
Курсовые работы по темам: линейная алгебра, введение в регрессию и статистические вычисления, экспериментальный дизайн, эконометрика, основы алгоритмов и вычислений.

НАВЫКИ

- Python
- SQL
- Машинное обучение
- Git
- Pandas
- Seaborn
- Scikit-learn
- NymPy

Рис. 6.1. Пример резюме начинающего дата-сайентиста

6.1.1. Структура

В этой части мы рассмотрим каждый раздел резюме, останавливаясь подробнее на том, что там следует написать.

> **ПРИМЕЧАНИЕ** Единственная цель резюме — убедить человека, просматривающего его по диагонали, что вас стоит пригласить на интервью.

РАЗДЕЛ: КОНТАКТНАЯ ИНФОРМАЦИЯ

САРА ДЖОНС

Сан-Франциско, Калифорния 534-241-6264
sarajones@gmail.com linkedin.com/in/sarajones sarajones.github.io github.com/sarajones

Не забудьте включить в резюме контактную информацию, чтобы рекрутер мог с вами связаться! Вы должны указать как минимум имя и фамилию, номер телефона и адрес электронной почты. Кроме того, можете дать ссылки на источники, где о вас можно больше узнать, включая ссылки на профили в социальных сетях, например LinkedIn, онлайн-базы кода, такие как GitHub, а также указать личные веб-сайты и блоги. Чтобы понять, что еще можно добавить, спросите себя: «Если кто-то это увидит, пойдет ли это мне в плюс?» Например, ссылка на портфолио вашего проекта — это восхитительно. А ссылка на профиль GitHub с копией учебного проекта — нет. Если вы выкладывали какую-нибудь из своих работ в открытый доступ, попробуйте найти способ отразить это в резюме.

Как правило, следует также указать город и штат проживания, чтобы рекрутер знал, как вы будете добираться до работы и придется ли вам переезжать ради нее. Некоторые компании не оплачивают переезд новых сотрудников, поэтому если вы живете далеко и не хотите рисковать, то можете не указывать свое местоположение.

Если ваша официально зарегистрированная фамилия и полное имя не совпадают с обычным именем, можете использовать последнее. В дальнейшем вам нужно будет сообщить работодателю свое официальное имя: это требуется для проверки данных, но вы не обязаны использовать официальное имя при отклике.

Наконец, откажитесь от адреса электронной почты, который может быть оскорбительным (например, i_hate_python@gmail.com) или адрес с истекшим сроком действия (например, ваш старый школьный имейл).

В этом разделе вы показываете, что обладаете достаточной квалификацией благодаря предыдущему опыту, стажировкам или курсам, которые прошли. Если

ваша предыдущая работа связана с Data Science, например вы занимались разработкой ПО, это прекрасно: выделите под нее значительную часть резюме. Если работа никак не связана с этой сферой, например вы были профессором истории искусств, укажите это, но не переусердствуйте. Сопроводите каждое место работы названием компании, месяцем и годом начала и окончания работы, названием должности и описанием своих обязанностей (минимум один пункт для обычных должностей и два или три для наиболее значимых). Если вы новичок или выпускник, можете указать стажировку и исследовательскую работу.

РАЗДЕЛ: ОПЫТ

ИЮНЬ 2019 — ПО НАСТОЯЩЕЕ ВРЕМЯ, САН-ФРАНЦИСКО
СТУДЕНТ ПРОГРАММЫ DATA SCIENCE, ПОТРЯСАЮЩИЙ БУТКЕМП

- создала веб-приложение на Python, которое рекомендует вам лучший район Нью-Йорка для проживания в зависимости от бюджета, образа жизни и работы;
- проанализировала 2200 статей на тему бизнеса из газеты New York Times, используя для этого обработку естественного языка, и визуализировала изменение тем с течением времени.

Этот раздел должен быть самым большим и потенциально может занимать половину резюме. Он также является самым важным, так как первое, на что рекрутеры будут обращать внимание, — это наличие у вас опыта работы с данными, который похож на требуемый в вакансии. Крайне важно оформить этот раздел правильно, поэтому мы подробно рассмотрим, как это сделать, в части 6.1.2.

РАЗДЕЛ: ОБРАЗОВАНИЕ

ОБРАЗОВАНИЕ

ИЮНЬ 2017, НОВЫЙ ОРЛЕАН
БАКАЛАВР ЭКОНОМИКИ, ВТОРОСТЕПЕННАЯ СПЕЦИАЛЬНОСТЬ — СТАТИСТИКА, КРУТОЙ УНИВЕРСИТЕТ
Средний балл успеваемости 3,65/4.
Курсовые работы по теме: линейная алгебра, введение в регрессию и статистические вычисления, экспериментальный дизайн, эконометрика, основы алгоритмов и вычислений.

В этом разделе вы рассказываете о своем образовании. В идеале вы должны показать, что обладаете набором навыков, которые будут полезны для работы в Data Science. Если вы обучались где-либо после старшей школы, укажите эти заведения, даже если у вас нет диплома, а также даты обучения (в том же формате, что и даты работы) и тематику обучения. Если вы ищете свою первую работу и у вас высокий (выше 3,3) средний балл, можете это указать. Если вы недавно закончили обучение и изучали статистику, математику, информатику или любые

другие предметы, в которых все это применялось (например, методы исследования социальных наук или инженерное дело), укажите их.

Рекрутерам будет очень интересно узнать, есть ли у вас специализация, имеющая отношение к сфере, например есть ли у вас степень в области Data Science, статистики, информатики или математики. Их также интересует уровень образования. Поскольку многие темы в этой области не рассматриваются до магистратуры, наличие диплома магистра засчитают вам в плюс. Рекрутерам обычно все равно, где вы учились, если только это не очень известный или престижный университет. Тем не менее если вы выпустились давно, то место вашего обучения не столь важно. При этом рекрутеры отмечают любые буткемпы, сертификаты или онлайн-программы, потому что они показывают, что вы прокачали свои навыки.

Хотя из раздела об образовании в резюме рекрутер может почерпнуть ценную информацию, вы никак не сможете улучшить этот раздел, если только не получите дополнительное образование или сертификат, о которых мы говорили в главе 3.

РАЗДЕЛ: НАВЫКИ

НАВЫКИ

- Python
- SQL
- Машинное обучение
- Git
- Pandas
- Seaborn
- Scikit-learn
- NymPy

В этом разделе вы можете перечислить все свои навыки в Data Science. В идеале рекрутер посмотрит сюда и скажет: «Отлично, подходит», — потому что эта информация соответствует вакансии. В этом разделе резюме перечисляются два типа навыков:

- *Программирование/работа с базами данных*. Это могут быть языки программирования, например Python и SQL, фреймворки и среды, например .NET или JVM, инструменты Tableau и Excel или экосистемы, например Azure и Amazon Web Services (AWS).
- *Методы анализа данных*. Второй тип — это методы анализа данных, такие как регрессии и нейронные сети. Возможно, вы владеете несколькими методами и все же постарайтесь сосредоточиться на ключевых, которые покажут, что у вас есть база плюс пара-тройка особенных навыков. Указав список вроде «регрессия, методы кластеризации, нейронные сети, анализ опросов», вы продемонстрируете, что знаете основы и более продвинутые темы.

Старайтесь не указывать более семи-восьми навыков, чтобы не перегружать людей. Также не стоит перечислять то, что не имеет никакого отношения к вакансии (например, малоизвестный язык программирования времен вашего обучения в магистратуре).

Перечислите только то, с чем вам будет удобно работать. Не указывайте язык, с которым не имели дела пять лет и который не хотите заново осваивать. Вас могут спросить обо всех пунктах из резюме. Если в вакансии требуются определенные навыки, а они у вас есть, обязательно укажите их! Ведь это именно то, чего ищут рекрутеры.

Мы не советуем использовать рейтинги, баллы или другие аналогичные способы, чтобы показать, насколько круто у вас развит каждый навык. Во-первых, рейтинги не значат ничего: любой может оценить себя на 5 из 5. Если вы поставите себе максимальные оценки, менеджер по персоналу может подумать, что вы врете или недостаточно самокритичны; если вы оцените себя ниже, в ваших способностях могут усомниться. Кроме того, непонятно, по каким критериям вы оцениваете каждый уровень. Что означает 5 из 5? Что вы считаете себя одним из лучших специалистов в мире, умеете решать сложные задачи или превосходите коллег в определенных навыках? Если HR действительно хочет, чтобы вы самостоятельно оценили свой уровень, он попросит вас об этом во время интервью.

Не перечисляйте «мягкие» навыки (soft skills), такие как критическое мышление и межличностную коммуникацию. Хоть они и играют важную роль в успешной работе, нет никакого смысла включать их в резюме, потому что они есть у всех. Если вы действительно хотите акцентировать на этом внимание, расскажите о том, как применили эти навыки в конкретных ситуациях. Кроме того, не нужно перечислять основные навыки, которые подразумеваются по умолчанию для этой вакансии, например не следует упоминать, что вы умеете работать с Microsoft Office Suite.

РАЗДЕЛ: DS-ПРОЕКТЫ (ДОПОЛНИТЕЛЬНО)

Если вы делали проекты с данными вне основной работы, можно выделить их в отдельный раздел. Он отлично подходит для тех кандидатов с небольшим опытом, которые занимались сторонними проектами в университете или буткемпе. По сути, вы говорите рекрутеру следующее: «Возможно, у меня не так много опыта работы, но это не важно, ведь я справился со всеми этапами процессов Data Science».

У каждого проекта должно быть название. Кроме того, вы должны описать, что и как делали, и рассказать о результатах. Проекты должны быть структурированными и содержательными, поэтому все, что описано в разделе 6.1.2 о создании контента, применимо и к ним. В идеале у вас должна быть ссылка на статью в блоге или хотя бы на репозиторий GitHub с информативным файлом README. Data Science — техническая область, где так легко показать проделанную работу, для чего и подходит этот раздел. Его можно пропустить при солидном опыте работы, но стоит все же рассказать о проектах в интервью.

РАЗДЕЛ: ПУБЛИКАЦИИ (ДОПОЛНИТЕЛЬНО)

Если вы публиковали статьи, связанные с Data Science, во время учебы в магистратуре или аспирантуре, перечислите их в резюме. Статьи в других областях, даже если это физика или вычислительная биология, тоже можно упомянуть, только кратко: так как они не связаны с Data Science напрямую, читающий вряд ли оценит что-то, кроме ваших стараний ради публикации. Вы можете описать проделанную во время исследования работу в разделе «Опыт работы», например указать, что был «создан алгоритм для анализа миллионов последовательностей РНК в минуту». Но сама по себе публикация, пусть даже в престижном журнале, о котором HR впервые слышит, вряд ли далеко вас продвинет.

ДРУГИЕ РАЗДЕЛЫ

Можно добавить и другие разделы, например «Награды», если вы выиграли соревнования Kaggle или получили стипендию или грант, но это не обязательно. Предоставлять рекомендации не нужно; позже такая возможность появится, если вы дойдете до этого этапа. Не стоит подробно расписывать свои объективные цели: это избыточная информация, учитывая все, что вы уже перечислили в резюме. Фраза «специалист по работе с данными, имеющий опыт работы с Python, ищет работу для развития навыков A/B-тестирования и моделирования» вряд ли приведет рекрутера в восторг!

СОЕДИНЯЕМ ВСЕ ВМЕСТЕ

Как правило, контактная информация размещается в начале, после чего следует наиболее важный раздел резюме. Если вы еще учитесь или недавно выпустились, поместите туда информацию об образовании; если у вас нет соответствующей работы или диплома, это место должен занять список DS-проектов; в противном случае поделитесь своим опытом работы. В разделах «Работа» и «Образование» укажите все места в обратном хронологическом порядке, начиная с последнего.

Мы видели множество эффективных форматов резюме. В случае с Data Science у вас есть немного свободы выбора в плане дизайна; какие-либо стандарты здесь в принципе отсутствуют. Несмотря на это, вы всегда должны стремиться к тому, чтобы ваше резюме можно было легко и быстро просмотреть по диагонали. Рекрутеры изучают резюме очень кратко: вы ведь не хотите, чтобы все это время было потрачено на поиск вашего последнего места работы? Не позволяйте дизайну отвлекать от содержания; подумайте, как он будет восприниматься читающим. Вот несколько полезных советов:

- Используйте понятные заголовки разделов, чтобы между ними можно было легко переключаться.

- Сделайте расстояние между абзацами больше, чтобы текст воспринимался легче.
- Выделяйте важные слова жирным шрифтом. Например, можно выделить должности, которые вы занимали в каждой компании.

Если эти идеи вам не подходят, возьмите шаблон из интернета или проконсультируйтесь со специалистом.

Как правило, лучше ограничиться одной страницей. Это делается по двум причинам: рекрутер крайне быстро просматривает ваше резюме, и за это короткое время он должен узнать о вас ровно то, что вы считаете наиболее важным. Кроме того, так вы показываете, что умеете емко доносить информацию и понимаете, какой опыт самый значимый. Если же человек отправляет резюме на 17 страницах (а мы видели и такое!), значит, он не имеет ни малейшего представления о том, почему подходит на должность; более того, он считает себя вправе отнимать столько времени у других людей.

> **ВЫЧИТКА** Очень важно проверить свое резюме! Несколько опечаток или грамматических ошибок могут отправить ваши старания в (метафорическое) мусорное ведро. Почему так жестко? Просматривая сотни резюме, рекрутеры выделяют два типа: замечательные (редко) и те, которые можно легко выкинуть. Второй тип требует некоторых правил отбора, и в дополнение к резюме кандидатов, которые явно не соответствуют требованиям, резюме с опечатками также легко станет причиной для отсеивания. Работа в Data Science требует внимания к деталям и проверки своих трудов; если вы не можете с этим справиться, когда выдаете все свои козыри в резюме, то что уж говорить о работе? Кроме проверки орфографии в текстовом редакторе попросите хотя бы еще одного человека внимательно прочитать ваше резюме.

Наконец, убедитесь, что резюме написано в одном стиле. Если вы сокращаете названия месяцев в разделе об образовании, сократите их и в разделе о работе. Хотя допускается использовать разные шрифты и кегль для заголовков и основного текста, все пункты должны быть одинакового формата. Пишите о предыдущем опыте в прошедшем времени, а для описания текущего места работы используйте настоящее. Так вы показываете, что обращаете внимание на мелкие детали и (опять же) помогаете рекрутеру быстро обрабатывать контент, так как ему не придется отвлекаться на разный шрифт или стиль. Одно несоответствие вряд ли будет стоить вам интервью, но иногда такие мелочи играют решающую роль: помните, что дьявол кроется в деталях.

6.1.2. Подробнее о разделе опыта: наполнение

Мы надеемся, что вы легко справитесь с датами и названиями в своей истории учебы и работы. Но как описать свой опыт (или DS-проекты), чтобы он выглядел ярче?

Распространенная ошибка, которую люди допускают при составлении резюме, — просто перечислить должностные обязанности: «Занимался отчетами для руководителей, используя SQL и Tableau» или «Преподавал математический анализ трем группам из 30 студентов». У такого подхода есть два существенных минуса. Во-первых, так в резюме вы указываете только то, за что отвечали, а не чего и каким образом достигли. Во-вторых, связь этих пунктов с Data Science может быть неочевидна. В двух предыдущих примерах вы могли бы описать то же самое, но другими словами: «Создание отчетов для руководителей по прогнозу продаж было автоматизировано с помощью Tableau и SQL, что позволило сэкономить четыре часа работы в неделю» или «Преподавал математический анализ для 90 студентов, средняя оценка учебного процесса учащимися составила 9,5/10 баллов, при этом 85 % учащихся сдали экзамен BC Calculus AP на 4 или 5 баллов».

Вы должны постараться максимально объяснить свой опыт с точки зрения навыков, которые можно перенести в Data Science. Даже если вы работали в другой сфере, приходилось ли вам иметь дело с данными? HR-менеджеры готовы рассматривать кандидатов без опыта в DS, но их нужно убедить. Если какая-либо из ваших работ может быть потенциально связана с получением данных и их обработкой, приложите максимум усилий и расскажите о процессе. Вы проанализировали 100 ГБ данных о звездах, когда получали степень кандидата наук по астрофизике? Собрали 30 файлов Excel, чтобы спланировать набор персонала для пекарни? Много где нужны данные для решения задач.

Вы использовали такие инструменты, как Google Analytics, Excel или Survey Monkey? Даже если они не понадобятся вам как дата-сайентисту, опыт работы с данными любого типа имеет значение. Какие коммуникативные навыки вы применяли? Приходилось ли объяснять технические или узкоспециализированные концепции во время кандидатского выступления или бизнес-доклада? Не отчаивайтесь, если не получается отыскать у себя навыки, применимые в DS: остальные советы по улучшению этих разделов все равно помогут. Но все же подумайте, как ваше образование или сторонние проекты могут продемонстрировать навыки работы с данными, особенно если у вас недостаточно опыта.

Если какие-то из ваших прежних мест работы имели мало общего с Data Science, и к тому же после вашего ухода оттуда прошло уже несколько лет, лучше ограничиться одним пунктом для описания каждого. Учтите, что не стоит исключать какие-то прошлые должности, если при этом в резюме образуется пробел в несколько месяцев. При смене многих компаний можно указать только три или четыре последних.

Возможно, вы заметите, что составить такой список намного проще для нынешней работы, чем для той, с которой вы ушли пять лет назад. Наш вам совет: начните вести список своих достижений и крупных проектов. Когда ежедневный прогресс не столь заметен, легко забыть, насколько впечатляет проделанная ра-

бота, если отступить и взглянуть на все в целом. Люди знают, что резюме — это не исчерпывающий перечень ваших заслуг, поэтому вряд ли кто-то подумает: «Ей потребовалось 15 месяцев, чтобы создать автоматизированную систему отслеживания и оценки потенциальных клиентов, которая сэкономила отделу продаж более 20 еженедельных часов работы, выполняемой вручную». Скорее, они подумают: «Ого, нам нужна подобная система!»

В целом основные пункты можно разделить на две категории. Первая — это большие достижения, такие как «Создана панель управления для отслеживания всех текущих экспериментальных программ и проведения расчетов статистической мощности». Вторая — это средние или итоговые результаты, например «Внедрено и проанализировано более 60 экспериментальных программ, что принесло компании дополнительный доход в размере более 30 миллионов долларов».

В любом случае каждый пункт должен начинаться с глагола и (в идеале) поддаваться количественной оценке. Вместо того чтобы писать «я проводил презентации для клиентов», напишите «создал более 20 презентаций для руководителей из списка Fortune 500». Будет даже лучше, если вы сможете количественно оценить свой вклад. Формулировка «провел 20 A/B-тестов для email-маркетинга, что привело к увеличению числа переходов на 35 % и росту продаж с атрибуцией в целом» намного эффективнее, чем просто «провел 20 A/B-тестов для email-маркетинга».

6.2. Сопроводительное письмо: основные положения

Цель резюме — предоставить HR-менеджерам факты о вашем опыте работы и образовании, а сопроводительное письмо необходимо, чтобы составить представление о вас как о личности. В нем можно объяснить, как вы нашли эту компанию, и подчеркнуть, почему вы идеально подходите. Если по резюме не всегда можно проследить ваш карьерный путь, то сопроводительное письмо отлично справляется с этой задачей. Даже если вы просто покажете, что знаете о направлении компании, заходили на ее веб-сайт или использовали ее продукт (если он доступен для физических лиц), это будет много значить. Сопроводительное письмо — ваш лучший инструмент, чтобы помочь рекрутерам прочитать ваше резюме между строк.

В отличие от резюме, сопроводительное письмо вовсе не обязательно. Но если вам известно, кому в организации его можно отправить, сделайте это; некоторые компании даже не рассматривают кандидатов без сопроводительного письма. Нередко они конкретно указывают, о чем можно написать, например это может быть ваша любимая методика обучения нейронной сети с учителем. Этим обычно проверяют, насколько внимательно кандидат прочел описание и требования к должности и не рассылает ли он шаблонные письма всем фирмам. Вы должны доказать работодателю, что умеете следовать инструкциям.

Распространенная ошибка, которую мы часто видим в сопроводительных письмах, — это чрезмерное внимание к тому, что компания может сделать для вас. Лучше не говорите: «Это было бы большим шагом в моей карьере». Задача рекрутера не в карьерной помощи всем желающим, а в найме полезных для компании сотрудников. Покажите им, что вы справитесь. Даже если эта работа станет для вас первой в Data Science, подумайте, какой схожий опыт у вас уже есть? Какие достижения (даже если они не связаны с данными) покажут, что вы много работаете и достигаете поставленных целей? Не следует себя недооценивать: хорошенько подумайте, чем можно заинтересовать рекрутера компании.

Как и резюме, сопроводительное письмо должно быть емким и занимать максимум одну страницу (а лучше три четверти). Сосредоточьтесь на своих сильных сторонах. Если в описании вакансии указаны четыре навыка, а вы преуспеваете в двух, сделайте акцент на них! Вы не обязаны извиняться за навыки, которых вам не хватает.

На рис. 6.2 показан пример сопроводительного письма.

6.2.1. Структура

Сопроводительные письма имеют не такой четкий набор правил по сравнению с резюме. Вместе с тем неплохо бы ориентироваться на эту хорошую общую структуру:

- *Приветствие.* В идеале необходимо узнать, как зовут нанимающего руководителя или рекрутера, который закрывает эту вакансию. Для начала поищите имя в самой вакансии. Даже если там указан только адрес электронной почты, Гугл может прийти на помощь. Проверьте LinkedIn и веб-сайт компании, чтобы сверить информацию.
- Даже если вы дойдете до самого вице-президента отдела, «перескочив через голову» (то есть сразу обратитесь к руководителю своего потенциального менеджера), вы все равно покажете, что подошли к вопросу серьезно. Если же выяснить имя все равно не получилось, направьте письмо менеджеру подразделения, который производит набор (например: «Уважаемый руководитель набора в отдел Data Science»). Избегайте обращения «Уважаемый господин или госпожа»: это выглядит архаично и шаблонно.
- *Вводный абзац.* Представьтесь, назовите интересующую вас должность и кратко объясните, почему вам интересны эта компания и сама работа. Если организация ведет блог по Data Science или кто-либо из ее специалистов по данным выступал с докладами или писал статьи о своей работе, упомяните здесь, что вы их смотрели или читали. Установите связь между тем, что вы узнали из этих презентаций, и тем, почему вам нравится эта должность или компания.

САРА ДЖОНС

Нью-Йорк, Нью-Йорк 534-241-6264
sarajones@gmail.com linkedin.com/in/sarajones sarajones.github.io github.com/sarajones

ПРИВЕТСТВИЕ	Здравствуйте, Джаред!
ВВОДНЫЙ АБЗАЦ	Меня очень заинтересовала вакансия дата-сайентиста в Потрясающей Компании. Я очень люблю читать блог Потрясающей Компании о Data Science с момента его запуска восемь месяцев назад. Решение, предлагаемое в статье об использовании тематического моделирования для автоматического генерирования тэгов, очень помогло мне в одном из моих собственных проектов, связанных с классификацией статей рубрики «Бизнес» газеты New York Times.
1–2 АБЗАЦА ОБ ОПЫТЕ РАБОТЫ В DATA SCIENCE	Недавно я закончила Потрясающий Буткемп, пройдя очный трехмесячный интенсив по Data Science. В Потрясающем Буткемпе я занималась разработкой, внедрением и реализацией DS-проектов на языке Python, включая сбор данных, их первичную обработку, машинное обучение и визуализацию. Для своего итогового проекта я собрала 3000 обзоров и оценок района от Районной Компании. Используя обработку естественного языка для обзоров и доступных списков из API Компании Недвижимости, я создала систему рекомендаций, которая способна подобрать район на основе вашего бюджета, предпочтений и текстового описания идеального для вас района. Испытать программу можно по ссылке моеклассноевеб-приложение.com До того как пойти в Потрясающий Буткемп, я работала консультантом по инвестициям в МегаКо. Когда я пришла в команду, все коллеги использовали для работы только Excel. Я превзошла поставленные передо мной цели, автоматизировав стандартные задачи с помощью Python: например, я автоматизировала составление еженедельного отчета о рыночных и отраслевых тенденциях, что позволило сэкономить несколько еженедельных часов работы моей команды. Я также разработала для них индивидуальную программу обучения Python. Инициатива была настолько успешной, что компания попросила меня разработать полный двухдневный семинар и отправила вести его в три других офиса, где в нем приняли участие более 70 консультантов.
ЗАКЛЮЧИТЕЛЬНЫЙ АБЗАЦ	Я уверена, что мои знания Python, образование в сфере экономики и статистики, а также опыт представления результатов коммерческой деятельности сделают меня отличным членом команды data science. Спасибо за Ваше время и внимание.
ПОДПИСЬ	С уважением, Сара Джонс

Рис. 6.2. Пример сопроводительного письма с выделенными составляющими

- *От одного до двух абзацев с примерами работы с данными*. Расскажите, как ваши предыдущие достижения связаны с этой должностью. Остановитесь подробнее на том, о чем вы упоминали в резюме, развернуто опишите одну должность или личный проект, приведите конкретные примеры. Следуйте принципу «не говори, а покажи результат»; вместо заявлений в духе «внимателен к деталям, организован и умею решать задачи» продемонстрируйте это на примере.
- *Заключительный абзац*. Поблагодарите рекрутера за уделенное время и внимание. Подведите итог своим навыкам, объяснив, почему вы подходите на эту должность.
- *Подпись*. «С уважением», «С наилучшими пожеланиями» и «Спасибо за рассмотрение моего резюме» — все эти варианты станут хорошим завершением. Не используйте разговорный или слишком неформальный стиль, например «Спасибо!» или «С самыми теплыми пожеланиями».

6.3. Адаптация

В двух предыдущих разделах изложены общие правила написания эффективного сопроводительного письма и резюме. Но лучший способ выделиться на фоне других кандидатов — это адаптировать их к каждой конкретной вакансии.

Вряд ли потенциальный руководитель станет первым, кто увидит ваш отклик; возможно, это будет даже не человек! В более крупных компаниях автоматизированные системы для подбора персонала проверяют резюме по ключевым словам. Такая система может посчитать, что «линейное моделирование» не соответствует требованию опыта работы с «регрессиями». Если резюме все же попадет в руки человеку, не факт, что он тоже примет правильное решение; может так получиться, что сотруднику отдела кадров не дали ничего, кроме списка обязанностей и инструкций по поиску перспективных кандидатов. Будет не очень здорово, если рекрутер так и не поймет, что проект с использованием «k-ближайших соседей» доказывает ваш опыт в кластерном анализе или что NLP — это аббревиатура для *обработки естественного языка*. Вам нужно, чтобы кто-то мог легко сравнить ваше резюме и вакансию и найти точные совпадения между вашим опытом и требованиями. Не перегружайте описание техническим жаргоном: лучше оставьте несколько ключевых слов (например, *R* или *Python*), чтобы резюме прошло отсев.

Мы рекомендуем составить «основное» резюме и сопроводительное письмо, откуда можно брать информацию по необходимости, а не писать все каждый раз с нуля. Такой метод особенно полезен при откликах на разные типы должностей. Если в одних вакансиях упор делается на машинное обучение, а в других — на исследовательский анализ, вам будет намного проще с готовыми пунктами и соответствующими ключевыми словами наготове. Основное резюме и сопрово-

дительное письмо могут занимать несколько страниц, но прежде, чем отправить их в компанию, убедитесь, что вы уложились в одну.

Адаптация отклика под вакансию вовсе не означает, что вам нужно включать один пункт или навык для каждого отдельного требования. Как мы уже говорили в главе 5, в описании вакансий перечисляются скорее пожелания к кандидату; попытайтесь выяснить, какие из них действительно нужны для работы. Иногда организации любезно делят навыки и опыт на «требования» и «пожелания», но даже если какая-то компания так не сделала, вы можете сами в этом разобраться по описанию обязанностей. Несмотря на желание работодателей получить кандидата с плюсиками по всем пунктам, большинство все же не рассчитывают на это всерьез.

Исключение составляют крупные технологические компании и известные быстрорастущие стартапы. Они набирают множество кандидатов и ищут причины для отказа. Такие организации очень беспокоятся о «ложноположительных» кандидатах, которые будут работать плохо или даже средне. А «ложноотрицательные» их вообще не волнуют: они не боятся пропустить какого-нибудь крутого специалиста, так как профи к ним и так выстраиваются в очередь. Для этих компаний вы обычно должны соответствовать если не на сто процентов, то уж точно на девяносто.

6.4. Реферальная программа

На веб-сайтах компаний и на портале по трудоустройству можно легко подать сохраненное резюме, нажав лишь одну кнопку. К сожалению, из-за такой простоты оно оказывается в стопке с сотнями или даже тысячами таких же «холодных откликов». Вот почему мы не рекомендуем так делать, пока вы не исчерпаете другие варианты. Поиск вакансий отлично помогает понять рынок, но лучший способ войти — попросить кого-нибудь придержать для вас дверь.

В большинство компаний можно попасть через потайной ход: с помощью *реферальной программы*. Это означает, что текущий сотрудник компании советует кого-то на должность, обычно путем подачи заявки и информации через специальную реферальную систему. Многие компании предлагают сотрудникам бонусы в виде пары тысяч долларов, если они приводят новых людей, которых в итоге берут в команду. Организациям нравятся работники, которые попали к ним таким образом, потому что они считаются прошедшими предварительную проверку: какой-то сотрудник компании, (предположительно) успешно справляющийся с работой, считает этого кандидата подходящим. Даже если кто-то официально вас не порекомендует, возможность написать в сопроводительном письме «я обсуждал эту должность с [имя звездного сотрудника X]», а также по-

просить этого человека, чтобы менеджер, проводящий набор, следил за вашим резюме, — уже огромный плюс.

Как найти людей, которые могут порекомендовать вас? Для начала загляните в LinkedIn: может, вы знаете какого-нибудь сотрудника заинтересовавшей вас компании. Даже если вы давно не общались с этим человеком, не бойтесь написать ему вежливое сообщение. Затем найдите людей, которые раньше с вами работали или учились. Вероятность получить ответ на сообщение повышается, если вы упомянете что-то общее. Наконец, поищите людей среди контактов второго уровня, чтобы узнать, что у вас общего. Если вы в хороших отношениях с кем-либо из ваших общих знакомых, обратитесь к этому человеку с просьбой вас представить.

Пишем эффективное сообщение

В статье «Do you have time for a quick chat?» («Есть минутка на поболтать?», http://mng.bz/YeaK) Трей Кози (Trey Causey), старший менеджер отдела Data Science в Indeed.com, дает несколько советов по эффективному обращению к незнакомому человеку с целью рассказать ему о своем проекте, поговорить о поиске работы или обсудить карьеру. Следуя этим рекомендациям, вы с большей вероятностью получите ответ, проведете продуктивную встречу и создадите хорошую основу для продолжительных взаимоотношений:

- Составьте список важных вопросов и идей, которые вы хотите обсудить, и включите его в электронное письмо.
- Предложите несколько вариантов времени (скажите, что разговор займет не более 30 минут) и мест для встречи рядом с работой человека.
- Купите ему обед или кофе.
- Приходите заранее.
- Задавайте конкретные вопросы и держите в уме цели разговора, соответствующие теме, которую вы указали в письме. Не просите «любой совет, который вы можете мне дать».
- Следите за временем и дайте знать, когда оно истекло; если человек захочет продолжить разговор, он это сделает.
- Поблагодарите своего собеседника и дайте обратную связь в будущем.

Вот такое простое сообщение составил Трэй:

«Привет, Трей. Я прочитал ваш пост в блоге об интервью на работу в Data Science. Я хотел бы пригласить вас на этой неделе выпить кофе в Storyville на Пайк-плэйс и задать вам несколько вопросов по этому материалу.

Я сейчас провожу интервью, и часть о вайтборд-кодинге меня очень заинтересовала. Я хотел бы узнать ваше мнение о том, как можно улучшить вопросы и ответы для вайтборд-кодинга, а также поделиться своим опытом.

Не могли бы вы уделить мне 30 минут, скажем, во вторник или среду на следующей неделе? Спасибо, что написали эту статью!»

При обращении к дата-сайентисту потратьте немного времени, чтобы узнать, чем он занимается. Ведет ли он блог или аккаунт в Twitter, есть ли у него репозиторий на GitHub, где он делится своей работой? Марк Мелун (Mark Meloon), специалист по данным в ServiceNow, написал в своем блоге статью «Climbing the relationship ladder to get a Data Science job» («Взбираясь по лестнице взаимоотношений ради работы в Data Science», http://mng.bz/O95o). В ней он рассказал, что наиболее эффективны сообщения с комплиментами по поводу опубликованного контента и просьбой ответить на несколько вопросов. Таким образом, вам не придется спрашивать о том, что они уже публично обсуждали, зато сможете сосредоточиться на получении совета, который не смогли бы найти в другом месте.

Помните, что вам могут помочь не только дата-сайентисты. Они могут лучше прочих рассказать, каково это работать в компании, но порекомендовать вас может любой сотрудник. Если кто-то из ваших знакомых работает в организации, куда вы хотите попасть, обратитесь к нему! Как минимум они могут дать вам представление о корпоративной культуре.

6.5. Интервью с Кристен Керер, инструктором по Data Science и создателем курсов

Кристен Керер (Kristen Kehrer) — инструктор по Data Science в Калифорнийском университете (подразделение Беркли), преподаватель на факультете Института менеджмента Emeritus и основатель Data Moves Me, LLC. Data Moves Me помогает командам по обработке и анализу данных передавать стейкхолдерам результаты моделей МО, чтобы предприниматели могли уверенно принимать решения. Кристен — магистр в области прикладной статистики и соавтор самиздата «Mothers of Data Science».

Сколько раз вы редактировали свое резюме?

Миллион! Я родилась в простой семье, мой отец был пожарным, а мама — домохозяйкой, поэтому меня не учили, как писать хорошее резюме. Но когда я заканчивала аспирантуру, то попросила помощи у более опытных коллег. А еще я всегда фиксировала все свои новые проекты или что-нибудь интересное, что можно было бы добавить в резюме. Я определенно не тот человек, который не обновлял резюме два года. Относительно недавно моя старая компания заплатила профессиональному карьерному консультанту, когда меня уволили. Мне прошлось узнать все о лучших методах составления резюме и эффективном позиционировании себя на рынке, чтобы получить отличную работу.

Я настоятельно рекомендую часто обновлять свое резюме. Иногда очень сложно сразу вспомнить все важные моменты, которые можно туда добавить, особенно если вы какое-то время работаете на одном месте. Например, я была соавтором

пары постеров в сфере здравоохранения, получивших награды. Это актуально не для каждой вакансии, на которую я откликаюсь, но если я рассматриваю должность в сфере здравоохранения, то хотела бы отметить это достижение. Если бы я ничего не записывала, то не смогла бы вспомнить ни соавторов, ни название постера.

Расскажите о самых распространенных ошибках среди кандидатов.
Их так много! До сих пор вспоминаю четырехстраничное резюме, в котором есть даже информация о том, что человек был тренером по плаванию. Еще одна ошибка — непонимание того, что системы отслеживания кандидатов плохо разбираются в определенных вещах. Если в резюме есть значки или диаграммы, многие старые автоматизированные системы могут их не распознать и все может закончиться тем, что вас автоматически забракуют. Мне также не нравится, когда люди ставят, скажем, три звезды за Python, потому что это не дает никаких подробностей. И когда какой-либо свой навык вы оцениваете на две звезды, то сообщаете, что это у вас получается плохо.

Вы адаптируете свое резюме к вакансии, на которую откликаетесь?
Я не зацикливаюсь на этом. Но почти все средние и крупные компании сейчас используют автоматизированную систему для подбора персонала, и вы должны иметь возможность получить высокий рейтинг с точки зрения соответствия ключевым словам. Если я вижу в описании что-то похожее на то, чем раньше занималась, но с другой формулировкой, то просто отредактирую пару слов, чтобы достичь совпадения с выражениями в вакансии.

Какие стратегии вы рекомендуете для описания опыта работы в резюме?
Я советую людям оптимизировать резюме для работы, которую они хотят, а не для той, которая у них есть. Вам не нужно составлять список всего, что вы когда-либо делали. Лучше подумайте, какой вклад вы сможете сделать в Data Science. Например, вы учитель математики, который объяснял технический или математический материал гуманитариям. Или, может быть, вы работали над проектом, и, хоть он никак не был связан с аналитикой, вам приходилось работать в нескольких командах и выполнять разные функции. В целом вы должны показать, что способны решать задачи, управлять собой, хорошо общаться и добиваться результатов. Наконец, с помощью сторонних проектов можно подчеркнуть свои технические навыки и проявленную инициативу.

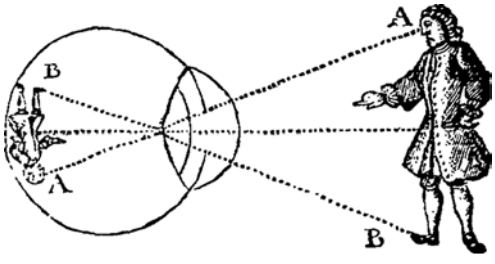
Ваш последний совет начинающим дата-сайентистам.
Начните уже откликаться на вакансии. Слишком много людей продолжают проходить онлайн-курсы, так как думают, что им нужно знать миллион вещей, чтобы стать специалистом. Но факт в том, что на работе вам все равно придется многое освоить. Я в этой сфере уже десять лет и все еще учусь. Откликнувшись, вы полу-

чите обратную связь. Если на ваше резюме никто не ответит, возможно, вы плохо себя позиционируете или просто не обладаете необходимыми навыками. Получите отзывы от нескольких человек, а затем выберите область, на которой нужно сосредоточиться, например возможность автоматизировать процессы в Python. Прокачайте этот навык, добавьте его в свое резюме и ищите следующие вакансии. Вы должны откликнуться, получить обратную связь, а затем все повторить и двигаться вперед, пока не получите работу.

Итоги

- Резюме — это не исчерпывающий список всех ваших предыдущих обязанностей. Оно необходимо для приглашения на интервью, а не для получения работы, поэтому постарайтесь, чтобы оно максимально соответствовало описанию вакансии.
- Сопроводительное письмо показывает, почему вы заинтересованы в компании и каким образом ваш опыт позволит внести ценный вклад.
- Общение с нынешними сотрудниками организации, особенно с дата-сайентистами, — лучший способ получить представление о специфике вакансий и корпоративной культуре.

7



Интервью: чего ожидать и что делать

В этой главе

- Что нужно интервьюерам.
- Стандартные вопросы.
- Соблюдение этикета общения с компанией.

Если задуматься над процессом интервью, то можно понять, насколько это сложно: каким-то образом вам нужно доказать совершенно незнакомым людям, что вы хорошо справитесь с работой, о которой судите только по нескольким абзацам в объявлении о вакансии. Вам могут задать технические вопросы любого уровня о различных технологиях, с которыми вы прежде не сталкивались. Кроме того, на интервью нужно будет получить достаточно информации о компании, чтобы понять, хотите ли вы вообще там работать. На все про все у вас есть всего несколько часов, при этом вы должны доказать, что являетесь профессионалом своего дела. Этого достаточно, чтобы вас прошиб холодный пот.

Хорошая новость в том, что при правильных подготовке и настрое интервью может превратиться в управляемый, терпимый и, возможно, даже приятный опыт.

В этой главе мы расскажем, что нужно интервьюерам и как настроиться на их волну. Обсудим технические и нетехнические вопросы, а также рассмотрим конкретный пример. А еще рассмотрим, как себя вести и какие вопросы задавать. Эти знания являются залогом хорошей подготовки к интервью.

7.1. Чего хотят компании?

Во время интервью компании ищут одного важного человека:

Кого-то, кто сможет выполнять эту работу.

Это единственное, что им нужно. Бизнес не ищет человека, который правильно отвечает на большинство вопросов на интервью, или кандидата с наибольшим количеством ученых степеней или опыта. Им нужен тот, кто может выполнять необходимую работу и помогать команде идти к поставленным целям.

Но что требуется для успешного выполнения работы? Вот несколько вариантов:

- *Наличие необходимых навыков*. Они могут быть как техническими, так и нетехническими. К техническим относятся навыки, которые мы рассмотрели в главе 1: некая комбинация математики и статистики, а также знания баз данных и программирования. Что касается нетехнической стороны, вам потребуется общая деловая хватка, а также умение управлять проектами и людьми, чувство вкуса для визуального дизайна и любые другие навыки, которые имеют отношение к должности.
- *Адекватное поведение*. Если вы говорите что-то оскорбительное, занимаете оборонительную позицию или у вас имеется ряд других недостатков характера, из-за которых другим людям будет сложно с вами сотрудничать, компания не захочет вас нанимать. Это означает, что во время интервью (а на самом деле всегда) следует быть приветливым, отзывчивым и позитивным. Речь не о том, что интервьюер должен захотеть пропустить с вами стаканчик. Это значит, что ваши будущие коллеги должны видеть в вас человека, с которым они хотят работать.
- *Способность выполнять работу*. Иметь соответствующие навыки недостаточно: их еще нужно уметь использовать! Вы должны знать, как находить решения и реализовывать их. В Data Science есть много мест, где можно застрять, например необходимость работать с неупорядоченными данными, решать сложные задачи, испытывать различные модели и приводить полученные результаты в порядок. Человек, который может справиться с каждой из этих задач, будет намного лучше выполнять свою работу, чем тот, кто сидит без дела в ожидании помощи, даже не попросив о ней. Перфекционисты тоже не идеальный вариант: их работа всегда находится на стадии шлифовки, а значит, вы рискуете не дождаться ее завершения.

Теперь вы знаете, какие именно качества компании ищут в кандидатах, и можете приступить к интервью. Мы построим наш рассказ об этом процессе вокруг трех основных идей.

7.1.1. Процесс интервью

В целом процесс интервью зависит от компании, но, как правило, оно проводится по стандартной схеме. Этот шаблон разработан для получения максимального количества информации о кандидате при минимальном количестве затраченного времени. Интервьюеры обычно очень загружены большим количеством собеседований и должны обеспечить справедливое сравнение кандидатов, поэтому процесс оптимизирован и последователен. Вот базовый план того, чего следует ожидать от интервью (рис. 7.1):

1. *Телефонное интервью*. Обычно оно занимает от 30 минут до часа и проводится специалистом по подбору технического персонала — человеком с большим опытом отбора кандидатов, который знает технические термины, но непосредственно сам не выполняет техническую работу. С точки зрения компании цель этого собеседования — проверить, есть ли у вас какие-либо шансы подойти на должность. Рекрутер должен отсеять людей, которые определенно не котируются, например тех, у кого нет соответствующей квалификации (или необходимых навыков), или грубиянов (которые не смогут эффективно работать с другими). С технической стороны интервьюер проверяет наличие у вас минимальных необходимых навыков, а не способность стать лучше всех. С гораздо большей вероятностью вас спросят: «А раньше вы использовали линейную регрессию?», чем «Как вы определите оценку максимального правдоподобия для гамма-распределения?» После первого собеседования по телефону иногда требуется еще одно с более технически подкованным специалистом. Если вы успешно преодолеете первый этап, то через несколько недель вас ждет...

Рис. 7.1. Четыре этапа интервью

2. *Интервью в офисе*. Оно часто длится от двух до шести часов и составляет основную часть процесса. Во время этого визита вы увидите будущее место

работы и познакомитесь с потенциальными коллегами. Это интервью дает компании возможность спросить вас о предыдущем опыте, навыках, а также ваших планах и ожиданиях в качестве дата-сайентиста. На этот раз с вами будут беседовать несколько человек, каждый из которых задаст вопросы на разные темы, включая нетехнические. Цель этого этапа — убедиться (с помощью технических вопросов), что у вас есть необходимые навыки и (с помощью поведенческих вопросов и того, как вы себя ведете) что с вами можно работать. Если это интервью пройдет хорошо, пора переходить к…

3. *Решению кейсов*. Вы получите описание реальной задачи и связанные с ней данные. Дома или непосредственно в офисе вы должны будете проанализировать данные, попытаться решить задачу и составить отчет. Затем вы должны представить его команде по найму. Это задание показывает, что у вас есть необходимые навыки (смотря насколько хорош отчет) и что вы справитесь с работой (смотря как много вы сделали). Такой этап есть не во всех компаниях; иногда его заменяют презентацией о вашей предыдущей работе. Если вы успешно справитесь и с этим этапом, готовьтесь к…
4. *Заключительному интервью с руководителем и офферу*. Его проводит старший менеджер, директор или другой руководитель группы. Цель этого этапа в том, чтобы вашу кандидатуру одобрил руководитель команды. Если вы дошли до него, значит, команда Data Science считает, что вы подходите на эту должность, поэтому это интервью редко отменяет ваши заслуги на предыдущих этапах. Обратите внимание, что оно часто проводится сразу после решения кейса, но бывает и так, что его проводят в начале или в конце первого собеседования в офисе. Если все пройдет хорошо, вы получите оффер в течение двух недель!

Если процесс отбора включает в себя все эти этапы, то оффер вам сделают где-то в промежутке между тремя неделями и двумя месяцами после того, как вы отправили свое резюме. Как видите, каждый этап предназначен для достижения различных целей компании.

В следующих разделах мы расскажем подробнее о каждом из них и научим, как правильно продемонстрировать свои навыки и способности в процессе.

7.2. Этап 1: первое телефонное интервью

Вашим первым взаимодействием с компанией, скорее всего, будет 30-минутный телефонный разговор с рекрутером. Важно произвести хорошее первое впечатление. Однако в зависимости от размера компании человек, с которым вы будете разговаривать, скорее всего, не будет иметь отношения к команде дата-сайентистов, поэтому *ваша цель — показать, что вы справитесь с обязанностями; не обязательно доказывать, что вы лучший из лучших*.

Почему? Задача рекрутера на этом этапе — отфильтровать неквалифицированных кандидатов. В разговоре он пытается оценить, стоит ли кому-то из отдела Data Science общаться с этим человеком. Часто люди откликаются на вакансии, не имея при этом должных навыков (или врут, что они у них есть), поэтому обязанность специалиста — не пропустить их дальше. Ваша цель — дать понять рекрутеру, что ваша квалификация достаточна для этой должности.

Скорее всего, вам зададут такие вопросы:

- *Расскажите о себе*. (Чисто грамматически это не вопрос, но относится к этой категории.) Здесь вас просят рассказать о себе за одну-две минуты. Рекрутер хочет услышать то, как вы описываете свой опыт, связанный с вакансией. Например, если вы претендуете на должность специалиста по принятию решений, от вас будут ждать рассказа о вашем академическом образовании и любых должностях или проектах, где вы занимались анализом. Важно, чтобы ваш ответ занимал от одной до двух минут. Если вы справитесь менее чем за минуту, сложится впечатление, что в вашей профессиональной жизни ничего особо не происходит. Если же рассказ будет длиться дольше, чем две минуты, то вы произведете впечатление человека, не умеющего выделять главное.
- *С какими технологиями вы знакомы*? Рекрутер проверяет, есть ли у вас технический опыт для выполнения работы. Помимо этого конкретного вопроса вы должны быть готовы к тому, что вас будут проверять на знание математики и статистики, баз данных и программирования, а также знания в области бизнеса (см. главу 1). Здесь стоит рассказать о любых известных вам технологиях, связанных с должностью или упомянутых в вакансии. Если вы не знаете чего-то из перечисленного (например, знаете Python, а не R), ничего страшного: просто скажите об этом честно и открыто. Если вы неплохо разбираетесь в стеке технологий, который использует компания (допустим, из разговоров с людьми, которые там работают), попробуйте выстроить свой ответ вокруг этого.
- *Чем вас заинтересовала эта должность*? Рекрутер пытается понять основную причину, по которой работа в компании вас привлекает. Хорошо продуманный ответ показывает, что вы подготовились и способны добиваться поставленных целей. Ответ вроде: «Я просто откликался на все вакансии в Data Science на LinkedIn» будет не в вашу пользу. Не отвечайте на такие вопросы слишком долго, просто покажите, что знаете, чем занимается компания, и искренне заинтересованы в этой должности. По возможности постарайтесь связать вакансию со своими опытом и интересами.

Хотя вас будут спрашивать и о вас самих, и о вашем опыте, у вас также есть шанс узнать больше о работе. Рекрутер должен потратить не менее 10 минут на то, чтобы рассказать о команде и должности, на которую вы претендуете. Задавайте

вопросы, чтобы убедиться, что эта работа придется вам по душе, а также чтобы продемонстрировать заинтересованность в ней. Спрашивайте о командировках, корпоративной культуре, о том, как меняется команда, какие у нее приоритеты и почему вообще открылась вакансия.

Возможно, что во время звонка рекрутер попытается узнать ваши ожидания по заработной плате. Вас могут спросить об этом прямо («На какую зарплату вы рассчитываете?») либо косвенно («Сколько вы сейчас зарабатываете?»). Плюс такого вопроса в том, что рекрутер заранее сравнивает предложение компании с вашими зарплатными ожиданиями, чтобы не тратить время на дальнейшее интервью с кандидатом, которого сумма не устроит. Но у этого вопроса есть и обратная сторона: рекрутер может попытаться установить зарплату ниже, чем компания способна предложить, ведь у него есть информация о ваших расценках. Постарайтесь избегать обсуждения зарплаты до более поздних этапов. Этому вопросу посвящена глава 8.

Во время разговора обязательно спросите о следующем этапе интервью и сроках. Рекрутер должен ответить что-то вроде: «Дальше будет еще одно собеседование: через неделю мы сообщим, собираемся ли вас пригласить». Абсолютно нормально спрашивать, чего стоит ждать дальше. Но не стоит интересоваться напрямую насчет приглашения на следующий этап. Это поставит рекрутера в затруднительное положение и, возможно, заставит чувствовать себя неловко, так как, скорее всего, он такие решения не принимает.

Если телефонное интервью пройдет успешно, вас пригласят в офис.

7.3. Этап 2: интервью в офисе

Этот этап — самое сердце процесса интервью. Компания пригласила вас в офис и выделила несколько часов на общение. Вы отпросились с работы, надели что-нибудь поприличнее и готовы встретить вопросы во всеоружии.

Цель этого интервью — помочь организации понять, сможете ли вы выполнять ту работу, для которой вас нанимают, и будете ли вы справляться с ней хорошо. В зависимости от типа компании и должности на собеседование может прийти от трех до десяти человек, при этом у каждого из них будут свои сильные и слабые стороны. Компания хочет найти лучшего среди них — или первого, кто окажется достаточно хорош.

На протяжении всего интервью продвигайте идею, что справитесь с работой на отлично. Цель не в том, чтобы оказаться самым умным, опытным или знающим больше всего технологий. Вместо этого вы должны быть кандидатом со здоровым балансом между адекватным подходом к работе, достаточными навыками для выполнения обязанностей и умением добиваться результатов.

Как происходит такое многочасовое интервью? Обычно оно включает следующее:

- Экскурсия по офису и знакомство с командой. Компания хочет дать вам понять, в каких условиях вы будете работать, а также впечатлить вас бесплатными напитками и закусками (если таковые имеются). На это уходит менее 15 минут, но мы советуем вам внимательно изучить обстановку, чтобы оценить уровень комфорта. Будет ли вам спокойно и легко на рабочем месте? Выглядят ли люди довольными и приветливыми? Не используются ли старые восьмилетние ноутбуки? В это же время с вами будет непринужденно беседовать кто-то из сотрудников компании. Будьте начеку! Это тоже часть интервью; если вы проявите себя как неприятный или злой собеседник, то вас могут не взять. Покажите себя с хорошей стороны, но, что важнее всего, будьте самим собой (если только ваше истинное «я» не придурок).
- Одно или несколько технических интервью. Эта часть может занять от 30 минут до нескольких часов в зависимости от требований компании. Вам будут задавать вопросы по многим темам, и, возможно, вам придется писать на доске или работать за компьютером. (Мы рассмотрим эти вопросы подробнее в разделе 7.3.1 и в приложении.) Цель этого этапа не в том, чтобы объяснить интервьюеру самые важные темы, которыми вы владеете, и даже не в том, чтобы показать ваше умение решать сложнейшие задачи. Вы должны доказать, что обладаете минимальным набором необходимых навыков для выполнения этой работы.
- Одно или несколько поведенческих интервью. Цель таких собеседований — понять, насколько хорошо вы ладите с другими и выполняете задачи. Вас будут много спрашивать об опыте работы, в том числе о том, как вы справлялись с трудными ситуациями и обеспечивали реализацию проектов. Вам могут задавать как общие вопросы («Расскажите мне о случае, когда вам приходилось взаимодействовать с проблемным коллегой»), так и более конкретные, связанные с Data Science («Как вы решили проблему неработающей модели и реализовали проект?»). Не факт, что ответить можно будет только правильно или нет; каждый ваш ответ может быть интерпретирован в зависимости от ситуации.

> **СОВЕТ** За день до интервью в офисе можно спросить у рекрутера, на что оно будет похоже. Как минимум вам должны сказать, с кем предстоит общаться и какие темы будут затронуты. Возможно, вас посвятят в некоторые детали того, чего следует ожидать во время технической и поведенческой частей собеседования. Эта информация поможет вам подготовиться.

Интервью в офисе эмоционально изматывают. Нужно уметь быстро переключаться с технической темы на вопросы о себе и своих мечтах, при этом надо произвести впечатление открытого человека, да еще и хорошо знающего свое

дело. В зависимости от размера компании с вами могут беседовать один или несколько сотрудников, и вы должны понравиться каждому из них. Один из лучших способов перестать нервничать в процессе — помнить, что интервьюеры тоже люди и так же, как и вы, хотят, чтобы у вас все получилось. Они ваши союзники, а не противники.

В следующих разделах мы более подробно рассмотрим этапы интервью. Ниже описываются особенности собеседований в офисе, а в приложении вы найдете развернутый список вопросов, которые вам могут задать, и примеры ответов, а также подробную информацию о том, как эти вопросы обдумывать.

Как одеться на интервью

Это один из наиболее часто обсуждаемых вопросов. Для работы в Data Science все усложняется тем, что вакансии могут быть в самых разных отраслях, в каждой из которых свои культура и дресс-код. То, что идеально подходит для одного интервью, может быть совершенно неуместным для другого.

При согласовании даты собеседования лучше всего спросить рекрутера, в чем обычно приходят соискатели, а также каков общий дресс-код компании. Этот человек заинтересован в вашем успехе, так что вряд ли собьет вас с правильного пути. В противном случае попробуйте пообщаться с кем-нибудь из сотрудников этой или схожей компании. Если других вариантов не осталось, логично предположить, что бюрократические отрасли (финансы, оборона, здравоохранение и так далее) имеют строгий дресс-код, тогда как в стартапах или крупных технологических компаниях обычно приходят в чем-то попроще. Избегайте крайностей (сандалии, шорты, коктейльные платья, шляпы-цилиндры); надевайте то, в чем вам удобно.

7.3.1. Техническое интервью

Многих дата-сайентистов техническое интервью пугает больше всего. Запросто можно вообразить, будто вы стоите у доски и понятия не имеете, как отвечать на заданный вопрос, осознавая при этом, что на работу вас в итоге не возьмут. (От одного только написания этого предложения у авторов уже задрожали коленки!)

Чтобы лучше понять, как пройти техническое интервью, вам стоит переосмыслить свое представление о нем. Если вы прочитали главу 4 и создали портфолио, считайте, что вы его уже прошли. Цель этого этапа —выяснить, есть ли у вас навыки, необходимые дата-сайентисту. А они у вас есть по определению, раз вы уже работали с данными! Если на собеседовании вас осуждают за то, что вы не знаете ответа на сложный вопрос, значит, это интервьюер не справился со своей задачей, а не вы. У вас за плечами есть необходимые навыки и опыт, и этот этап как раз создан для того, чтобы это доказать. Если интервью не позволяет этого сделать, это не ваша вина.

На этом этапе вы показываете интервьюеру, что у *вас есть необходимые для работы навыки*. Продемонстрировать их наличие — это не то же самое, что просто правильно отвечать на каждый заданный вопрос. Человек может отвечать так, как хочет интервьюер, и при этом провалить собеседование, а может отвечать совершенно неправильно и при этом успешно его пройти. Рассмотрим два варианта ответа:

> *Интервьюер*: Что такое *k*-фолд перекрестная проверка?
> *Ответ А*: Вы случайным образом разбиваете данные на *k* четных групп и используете их в качестве тестовых данных для *k* моделей.
> *Ответ Б*: Вы случайным образом делаете выборку и используете ее в качестве тестовых данных для модели *k* раз. Затем берете среднее значение моделей и используете его. Этот метод на самом деле является методом работы с переобучением, потому что у вас есть набор моделей, каждая из которых имеет разный набор данных для обучения. Я использовал этот метод в основном проекте по прогнозированию стоимости жилья, который указан в портфолио.

Технически ответ А является правильным, а ответ Б — нет (хотя кандидат и описал перекрестную проверку, технически она не является *k*-кратной, потому что данные не были разделены на четные группы). При этом ответ А не дал интервьюеру никакой информации, кроме того, что интервьюируемый знал определение, в то время как ответ Б показал, что кандидат знал этот термин, понимал, почему он был использован, и имел опыт работы с ним. Этот пример показывает, почему важно прежде всего думать о том, как донести свою компетенцию до рекрутера.

В частности, во время технического интервью вы можете руководствоваться следующими советами, чтобы звучать убедительнее:

- *Объясните ход ваших мыслей*. Постарайтесь не отвечать коротко; объясните, как вы получили решение. Это покажет интервьюеру, как вы рассуждаете, и докажет, что вы на правильном пути, даже если ответ неверный. Дадим одно предостережение: хотя повторение вопроса вслух может показаться полезным (например, «Хм... сработает ли здесь линейная регрессия?»), некоторые интервьюеры могут воспринять такое поведение как признак незнания. Попрактикуйтесь отвечать на вопросы и с самого начала учитесь рассуждать.
- *Ссылайтесь на свой опыт*. С помощью рассказов о своих проектах или проделанной работе вы постоянно подкрепляете разговор фактами о реальных практических навыках. Такой подход сделает ваш ответ убедительнее, либо же у вас появится альтернативная тема для обсуждения, если вы ушли от изначальной. Только не переусердствуйте: если вы слишком много говорите о своем опыте вместо того, чтобы отвечать на заданный вопрос, может сложиться впечатление, что вы уходите от ответа.

- *Если не знаете ответа, скажите об этом честно*. Вполне возможно (и нормально!) не знать ответов на все вопросы. Постарайтесь не врать и говорите то, что знаете. Если, например, вас спросят, что такое полусоединение, а вы понятия не имеете, то можете сказать что-то вроде: «Я не слышал раньше о таком, но подозреваю, что это может быть связано с внутренним соединением». Лучше открыто говорить о том, чего вы не знаете, чем уверенно ошибаться; интервьюеры часто настороженно относятся к людям, которые не в курсе, каких именно знаний им не хватает.

> **СОВЕТ** Во время интервью может сработать инстинкт отвечать как можно быстрее. Постарайтесь побороть его; лучше дольше подумать, но ответить хорошо, чем ответить быстро, но плохо. Из-за стресса во время собеседования темп речи ускоряется, поэтому потренируйтесь отвечать на вопросы заранее, чтобы чувствовать себя комфортнее.

Ниже приведены общие типы вопросов, которые обычно задают на техническом интервью. Опять же, советуем ознакомиться с приложением.

- *Математика и статистика*. Эти вопросы проверяют, насколько хорошо вы понимаете теорию, которая является необходимой основой для работы в Data Science. К ним относятся:
  - *Машинное обучение*. Эта тема включает в себя знание различных алгоритмов машинного обучения (*k*-среднее, линейная регрессия, случайный лес, анализ основных компонентов, вспомогательные векторные машины), различных методов использования алгоритмов машинного обучения (перекрестная проверка, бустинг) и общий опыт их практического применения (например, ситуации, когда определенные алгоритмы не работают).
  - *Статистика*. Вам могут задать чисто статистические вопросы, особенно если вы работаете в той области, которая с ней связана, например экспериментирование. Эти вопросы могут включать статистические тесты (например, t-тесты), определения терминов (например, ANOVA и *p*-значение) и вопросы о распределениях вероятностей (например, поиск среднего значения экспоненциальной случайной переменной).
  - *Комбинаторика*. Эта область математики охватывает все, что связано со счетом. К логическим задачам относятся такие вопросы, как: «В сумке есть шесть разноцветных шариков. Сколько их комбинаций останется, если вы вытащите два шарика без замены?» Эти вопросы не имеют ничего общего с работой дата-сайентиста, но интервьюеры порой считают, что так они смогут больше узнать о ваших навыках решения задач.

- *Базы данных и программирование*. Эти вопросы проверяют, насколько эффективно вы будете выполнять компьютерную часть работы по анализу данных. Она включает:
  - *SQL*. Практически на любом интервью в сфере Data Science вам будут задавать вопросы по базам данных SQL. Эти знания необходимы в большинстве случаев, и если вы знаете SQL, то сможете быстрее приступить к работе. Будьте готовы к тому, что вас спросят, как писать SQL-запросы для выборочных данных. Например, вам могут дать таблицу оценок учащихся в нескольких классах и попросить найти имена тех, кто набрал наибольшее количество баллов в каждом классе.
  - *R/Python*. В зависимости от компании вас могут попросить ответить на общие вопросы по программированию, написать псевдокод или решить конкретные задачи с помощью R или Python (выберут тот язык, который используется в организации). Не переживайте, если вы знаете R, а в компании принято писать на Python (или наоборот); как правило, всему можно научить, и компании с радостью делают это. Будьте готовы к вопросу, связанному с написанием кода (например: «Как бы вы отфильтровали таблицу в R/Python, чтобы она включала только строки выше 75-го процентиля столбца оценок?»).
- *Опыт в сфере бизнеса*. Эти вопросы во многом зависят от компании, в которую вы устраиваетесь. Здесь организация хочет понять, насколько вы знакомы с ее деятельностью. Хотя вы всему можете обучиться в процессе работы, работодатель предпочел бы, чтобы вы уже все знали и умели. Вот несколько примеров вопросов из различных сфер:
  - Компания электронной торговли. Каков показатель CTR электронного письма? Как он соотносится с показателем просмотров и как их можно рассмотреть иначе?
  - Логистика. Как оптимизировать производственные очереди? Что следует учитывать при управлении заводом?
  - Некоммерческая организация. Как НКО можно определить рост спонсоров? Как выяснить, что большое количество спонсоров не продлит контракты?
- *Логические задачи с подвохом*. Помимо задач, связанных непосредственно с Data Science, вам могут задать и общие вопросы на логику и находчивость. Считается, что их цель — проверить ваш интеллект и способность импровизировать. На практике ничего подобного не происходит. Google провел масштабное исследование (http://mng.bz/G4PR), в ходе которого было выявлено, что такие вопросы не позволяют спрогнозировать, как кандидат будет выполнять свою работу; они нужны только для того, чтобы интервьюер почувствовал себя

умным. Эти вопросы звучат примерно так: «Сколько бутылочек шампуня есть во всех отелях в США?» Обычно, загуглив, можно узнать, задают ли такие вопросы в крупных компаниях (а также выяснить, какие именно).

Сложно сказать, что конкретно у вас спросят и сколько времени дадут на ответ; эти факторы в основном зависят от компании и человека, который проводит интервью. Постарайтесь сохранять спокойствие и уверенность, даже если вы не можете найти ответ на какой-то вопрос. Если интервьюер что-то вам говорит, когда вы ответили частично, возможно, он считает, что все идет хорошо, и старается подтолкнуть вас в нужном направлении. Часто вопросы на интервью могут охватывать такое количество тем, что ни один дата-сайентист не сможет ответить на все из них, поэтому предполагается, что и вы оставите часть из них без ответа.

Вопросы интервьюеру

Каждый раз под конец интервью в офисе спрашивают: «У вас есть вопросы?» Это одна из ваших единственных возможностей получить правдивую информацию о работе, поэтому используйте ее с умом! Вы можете узнать больше об используемых технологиях и задачах, а также о том, как работает команда. Ваши вопросы показывают, что вы искренне заинтересованы в компании, поэтому продумайте их заранее. Вот несколько примеров:

- *«Какие технологии вы используете и как обучаете новых сотрудников?»* Это замечательный вопрос, если во время собеседования вам не рассказали подробно о стеке технологий. Ответ даст вам представление о том, есть ли в компании формальные процессы обучения или компания ждет, что сотрудники научатся всему сами.
- *«Кто стейкхолдер нашей команды и каковы взаимоотношения с ними?»* Вы спрашиваете, кто будет руководить работой. Если взаимоотношения плохие, вы можете стать заложником требований стейкхолдера даже против своей воли.
- *«Как проводится контроль качества работы в отделе Data Science?»* Поскольку команда не должна сдавать работу с ошибками, процесс должен предусматривать проверку. На практике работа многих команд по анализу данных не проверяется, и тогда они обвиняют работодателя, если что-то идет не так. Избегайте такой токсичной атмосферы!

7.3.2. Поведенческое интервью

Поведенческое интервью проводится для проверки вашего стиля общения и помогает команде дата-сайентистов лучшее понять вас и ваш опыт. В то время как на технические вопросы отводится один или два временны́х блока, поведенческие могут задавать на протяжении всего вашего нахождения в офисе. Вы можете столкнуться с ними во время часового собеседования с HR-ом, в заключительные

десять минут технической части или даже болтая с одним из сотрудников в ожидании кого-то еще. Так что будьте готовы ответить в любое время.

Вот несколько примеров распространенных вопросов на интервью, которые можно ожидать. Больше поведенческих вопросов и ответов представлено в разделе А.4 приложения:

- «*Расскажите о себе*». Этот «вопрос» задается во время телефонного интервью. Вы можете слышать его каждый раз, когда беседуете с новым человеком. Опять же, попробуйте дать краткий ответ, который займет от одной до двух минут, но на этот раз адаптируйте его под конкретного собеседника.
- «*Над каким проектом вы работали и чему научились?*» Этот вопрос призван показать, сможете ли вы вспомнить свой прошлый проект и проанализировать его. Вы смогли определить, что у вас получилось хорошо, а что пошло не так?
- «*Каков ваш главный недостаток?*» Этот вопрос выводит из себя, потому что с точки зрения теории игр нужно дать ответ, который выдает как можно меньше ваших слабых мест. На практике рекрутер пытается проверить, осознаете ли вы собственные слабые стороны и есть ли сферы, в которых вы активно пытаетесь совершенствоваться.

Обратите внимание, что все эти вопросы открытые и для них нет правильных ответов. Но некоторые способы выразить мысль могут значительно улучшить впечатление аудитории.

Ваши ответы на большинство из этих вопросов, особенно на те, которые касаются вашего опыта, должны соответствовать общей схеме:

1. Объясните вопрос своими словами, чтобы показать, что вы его поняли.
2. Объясните особенности ситуации, в которой возникла проблема, сосредоточив внимание на ее причинах.
3. Опишите действие, которое вы предприняли для решения проблемы, а также полученный результат.
4. Подведите итог о полученном опыте.

Рассмотрим следующий вопрос: «Расскажите мне о случае, когда вы сделали что-то для стейкхолдера и получили отрицательный результат». Ответом может быть нечто подобное: «То есть вы спрашиваете, когда я разочаровал кого-то своей работой? [*1. Пояснение вопроса*] Это произошло в предыдущей компании, когда я должен был составить отчет о росте числа клиентов. Наша команда была загружена множеством разных задач, поэтому у меня было мало времени, чтобы сосредоточиться на запросе, поступившем от одного директора. Когда я предоставил отчет, на который потратил всего один день, директор был очень разочарован [*2. Описание проблемы*]. Во-первых, я извинился за то, что не оправдал ожиданий; затем мы вместе постарались понять, как можно упростить

требования и получить при этом нужный результат [*3. Предоставление решения*]. Из этого опыта я понял, что лучше всего заранее сказать, что вы не можете удовлетворить чью-либо просьбу, чтобы прийти к решению, которое устроит всех [*4. Какой урок вы извлекли*]».

Вопросы поведенческого интервью хороши тем, что все они похожи, поэтому можно заранее подготовиться! Если у вас есть три или четыре случая, когда вы работали в сложных условиях, с трудными коллегами и разбирались с собственными ошибками, можно опираться на них при большинстве ответов. Это гораздо проще, чем пытаться придумать историю и импровизировать на ходу. Если у вас есть время, можете потренироваться рассказывать вслух об этих случаях другу, чтобы научиться лучше структурировать свою речь.

Практически на каждом интервью вас будут просить описать выполненный ранее проект, поэтому стоит хорошо подготовиться. Идеальный для этих целей проект должен включать моменты, перечисленные выше: сложная ситуация, в которой вы преодолевали трудности, особенно взаимодействие с проблемными коллегами, а затем нашли решение. В идеале история должна быть выдержана в четыре этапа.

Часто бывает сложно подобрать проект, который подходил бы по всем этим параметрам, особенно если вы начинающий дата-сайентист. Если вы понимаете, что не можете найти подходящий ответ, попробуйте просто рассказать о проекте из портфолио. Даже такой ответ, как: «Я подумал, что было бы здорово проанализировать необычный датасет, поэтому я извлек данные, обработал их и получил интересный результат, который описал в блоге», показывает интервьюеру, что вы умеете выполнять анализ.

Наилучшая методика ответа на поведенческие вопросы уже существует и основана на опыте огромного количества людей. Вы можете продумать все вплоть до точной формулировки и количества секунд, которые потратите на каждый ответ. При этом большинство техник не исключительны для Data Science, так что о них довольно просто узнать из книг и статей, посвященных интервью. Мы приводим свои рекомендации в разделе полезных материалов после главы 8.

7.4. Этап 3: решение кейса

Если вы хорошо прошли интервью в офисе, вам предложат решить кейс: то есть реализовать небольшой проект, который покажет работодателю, насколько хорошо вы справляетесь с задачами на практике. Кто-то из специалистов команды предоставит вам набор данных, нечеткую задачу и определенное время для ее решения. Вас могут попросить справиться с кейсом во время офисного интервью, выделив для этого час или два, или же дать больше времени, например выходные, чтобы сделать

это дома. Обычно компания разрешает использовать языки программирования или инструменты, с которыми кандидат хорошо знаком, хотя не исключено, что вас могут ограничить стеком технологий, которые используются непосредственно в компании. По окончании времени следует коротко представить результат и обсудить его с группой людей из команды Data Science. Вот несколько примеров кейсов:

- На основании данных о рекламных рассылках фирмы и размещенных заказах определите, какая из промокампаний показала наилучшие результаты и как организации следует изменить маркетинговую стратегию в будущем.
- На основании текста 20 000 твитов, в которых упоминается компания, сгруппируйте твиты по темам, которые, по вашему мнению, будут полезны для команды маркетинга.
- На сайте компании было проведено дорогостоящее A/B-тестирование, но в процессе сбор данных прекратился. Возьмите данные эксперимента и посмотрите, можно ли из них извлечь какую-либо ценную информацию.

Обратите внимание, что каждый кейс, приведенный в примере, не связан с Data Science напрямую. Такие вопросы, как: «Результат какой кампании оказался лучше?» целесообразны в контексте бизнеса, но алгоритм с подобным вопросом вы применить не сможете. Эти кейсы полезны в качестве инструментов собеседования, поскольку они требуют, чтобы вы прошли путь от самого начала задачи до ее решения.

Так что же именно компания хочет получить в результате? Работодателю важно следующее:

- *Можете ли вы взять неопределенную и нерешенную задачу и придумать методы ее решения*? Вполне возможно, что у вас не получится решить эту задачу, но если в целом вы мыслите в правильном направлении, то доказываете, что обладаете техническими навыками и можете добиваться целей.
- *Можете ли вы работать с неупорядоченными реальными данными*? Предоставленные вам данные, скорее всего, потребуют фильтрации, группировки, конструирования признаков и обработки недостающих элементов. Давая вам сложный датасет, компания ставит задачу, подобную тем, что вам придется решать на должности.
- *Можете ли вы структурировать анализ*? Компания хочет знать, умеете ли вы рассматривать данные методично и обдуманно, а не начинаете заниматься вещами, не имеющими отношения к задаче.
- *Можете ли вы составить содержательный отчет*? Вам нужно будет рассказать о своей работе и, возможно, сформировать такие документы, как Jupyter Notebooks или отчеты на R Markdown. Компания хочет знать о ваших возможностях сделать что-то полезное для бизнеса и умении структурировать информацию.

Хорошая новость в том, что навыки и методы, необходимые для решения кейсов, те же, что и для создания отличного портфолио: нужно взять данные, задать вопрос и получить результат. Еще лучше, если вы практиковались в ведении блога: это имитирует рассказ для кейса на интервью!

Вы должны принять во внимание несколько незначительных различий между портфолио и практическим кейсом:

- На решение кейса вам дается ограниченное время. Оно может измеряться по календарю; например, вам могут выделить неделю со дня получения материалов, или количеством отработанных часов, скажем не более двенадцати. Такой короткий промежуток означает, что вам придется выработать стратегию, чтобы уложиться в дедлайн. В целом этапы очистки и подготовки данных занимают гораздо больше времени, чем ожидают дата-сайентисты. На обычное объединение таблиц, фильтрацию искаженных символов из строк, а также загрузку данных в среду разработки могут уйти долгие часы, и, как правило, такой работой компанию не впечатлить. Постарайтесь не тратить слишком много времени на приведение данных в идеальное состояние, иначе вы рискуете не успеть с анализом.
- При решении кейса вас оценивают по презентации, поэтому она должна быть подготовлена на высшем уровне и содержать интересные результаты. Сам процесс создания презентации может показаться неинтересным и менее важным, чем анализ, поэтому многие дата-сайентисты откладывают его на самый конец. Такая практика — не лучшая идея, потому что вы можете не вписаться в дедлайн или понять, что анализ оказался не таким интересным, как вы предполагали, но у вас уже не будет времени на внесение изменений. Начните работать над презентацией как можно раньше и дополняйте ее по мере работы над анализом.
- Еще одно отличие кейса заключается в том, что у вас очень специфическая аудитория: небольшое количество людей, которым вы рассказываете о проделанной работе. Вы никогда наверняка не знаете, кто будет просматривать ваше портфолио, но с кейсами все гораздо проще; можно сделать анализ весьма узконаправленным. По возможности узнайте, кому вы будете представлять результаты кейса. Если среди аудитории будут только дата-сайентисты, можно сделать уклон в техническую сферу, например добавить подробные сведения об использованных методах машинного обучения и объяснить свой выбор. Если вас будут слушать люди, связанные с бизнесом, постарайтесь не слишком углубляться в технические детали и сосредоточьтесь на том, как сделанные выводы повлияют на бизнес-решения. Если аудитория представляет собой сочетание специалистов по данным и стейкхолдеров из бизнеса, постарайтесь сбалансировать ответ, чтобы у вас было достаточно аргументов для обеих сторон.

Сама презентация обычно занимает от 20 до 30 минут, а 10–15 минут в конце отводится на вопросы аудитории. Попрактикуйтесь в своем выступлении и распланируйте, что вы хотите сказать в каждой его части. Практика также помогает соблюдать регламент; пять минут для презентации явно недостаточно, но и на пятьдесят ее не стоит растягивать. По окончании вас обязательно завалят вопросами на самые разные темы. Сначала вы можете отвечать про параметр в модели, а затем объяснять, как результат вашей работы влияет на бизнес. Перед ответом обдумайте вопрос, чтобы лучше сориентироваться. Если у вас нет ответа, в котором вы уверены, лучше всего сказать: «Я точно не знаю, но...», а затем предложить несколько вариантов решения. Если возможно, добавьте что-то, в чем точно хорошо разбираетесь.

7.5. Этап 4: итоговое интервью

Когда решение кейса будет завершено, вероятно, в тот же день с вами проведут последнее интервью. Это будет сотрудник, который примет окончательное решение, например руководитель группы по Data Science или технический директор. В зависимости от того, как в компании устроен процесс интервью, не факт, что этот человек будет знать ваши результаты с прошлых этапов. Цель этого собеседования в том, чтобы он одобрил вашу кандидатуру.

Здесь сложно сказать, какие вопросы вам зададут на этом этапе, так как во многом это зависит от собеседника. ИТ-специалист может расспрашивать о техническом опыте и навыках, а бизнесмен — о ваших подходах к решению задач. Независимо от того, кто будет проводить интервью, вам определенно следует быть готовым к вопросам, похожим на поведенческие, например: «Как вы справляетесь с трудными ситуациями и проблемами?» Отвечая на эти вопросы, крайне важно быть открытым и искренним.

7.6. Оффер

Если все пройдет хорошо, в течение недели или двух после последнего интервью вам позвонит кто-то из компании и сообщит, что вам готовы сделать оффер. Поздравляем!

В следующей главе мы подробнее рассмотрим, что можно считать хорошим оффером, как сравнивать его детали для различных вакансий в Data Science и как попросить компанию улучшить его условия.

К сожалению, разочарования тоже случаются; оффер можно не получить. Немного погоревав о потере потенциальной работы, посмотрите на эту ситуацию как на возможность узнать, что следует прокачать к следующему интервью. Если

вы прошли только первое телефонный этап, это говорит о том, что ваша базовая квалификация не подходит для выбранной должности. В этом случае получше присмотритесь к вакансиям, на которые откликаетесь. Если вы дошли до интервью в офисе или до решения кейса, но дальше не получилось, вероятно, у компании есть конкретная причина, по которой вы не подошли. Попытайтесь понять, чему следует уделить внимание в следующий раз. Если вы прошли последнее интервью, но оффер не получили, обычно это означает, что вы хорошо подходите для этой должности, но кто-то другой вас все же обошел. В этом случае ничего не поделаешь; продолжайте откликаться на аналогичные вакансии. Не следует обращаться в компанию, чтобы выяснить причину отказа: вы вряд ли получите честный ответ, а вопрос будет сочтен непрофессиональным.

Обратная связь

После каждого этапа интервью у вас может возникнуть желание связаться с кем-то из компании в той или иной форме. Так можно выразить благодарность людям, с которыми вы познакомились, а также получить больше информации о том, как идут дела. Однако если вы сделаете это неправильно, то есть риск показаться навязчивым или отчаявшимся и тем самым поставить под угрозу ваши шансы получить работу. Вы можете использовать любой из трех вариантов в зависимости от того, на каком этапе находитесь:

- *Прежде чем кто-либо из компании свяжется с вами.* Если вы откликнулись на вакансию, но не получили ответа, не стоит звонить и писать. Отсутствие ответа — признак незаинтересованности работодателя.
- *После взаимодействия, но до того, как вы встретились с кем-либо лично.* После телефонного интервью вам следует напомнить о себе лишь в том случае, если вы не уверены в своем статусе в процессе. Если прошел указанный срок, в который вам должны были сообщить о дальнейших действиях, допустимо отправить один имейл. В этом случае просто попросите, чтобы вас держали в курсе дела.
- *После личной встречи*. Вы можете (но это не обязательно) отправить короткое письмо с благодарностью тем, кто проводил интервью. Если вам не ответят в обозначенный изначально срок, можно также написать рекрутеру с просьбой сообщить результат.

7.7. Интервью с Райаном Уильямсом, старшим специалистом по принятию решений в Starbucks

Райан Уильямс (Ryan Williams) недавно покинул должность DS-менеджера в фирме, занимающейся консалтингом в области маркетинга и продаж, где он руководил процессом приема на работу дата-сайентистов. Сейчас в составе команды ана-

литиков компании Starbucks он помогает консультировать людей, принимающих решения по программе Starbucks Rewards. У него есть степень бакалавра, которую он получил в Вашингтонском университете по двум направлениям: статистика и экономика. До Starbucks он работал в сфере консалтинга.

Что нужно сделать, чтобы «попасть в яблочко» на интервью?

Главное — это подготовка. На интервью вам потребуется целый набор специфических навыков. Многие думают, что просто войдут в кабинет, где их опыт проявится сам собой. Я сам так думал и видел, как другие кандидаты приходили на интервью с теми же мыслями. Но на вопрос вроде: «Расскажите нам о случае, когда вы столкнулись с проблемами в общении» ответить с ходу не так-то просто, и вы можете ходить вокруг да около, бурча что-то в процессе. Для прохождения интервью действительно нужны кое-какие навыки, а подготовиться можно, просто прочитав стандартные вопросы, с которыми вам предстоит столкнуться. Независимо от компании вам зададут некоторые поведенческие и технические вопросы, а также такие, которые связаны с бизнес-процессами.

Поэтому лучше знать общие поведенческие вопросы. Ознакомьтесь с типами потенциальных технических вопросов и бизнес-кейсов. Если вы не подготовитесь заранее, чтобы правильно продемонстрировать свой опыт, не факт, что вы получите такую возможность на интервью.

Как вы поступаете в случаях, когда не знаете ответа?

Помню один случай: я проходил собеседование в одной из крупных технологических компаний, и мне задавали все более и более сложные вопросы по статистике. Дошло до того, что меня спросили о чем-то совсем уж научном. Кажется, это было связано с определенной функцией распределения вероятностей. Меня спросили, как можно использовать ее производящую функцию моментов, чтобы найти эксцесс кривой распределения. Я сказал что-то вроде: «Ну… возможно, я смог бы ответить, когда учился в колледже, но сейчас точно не смогу».

В общем, мой ответ явно разочаровал интервьюера. Я мог бы начать возмущаться по этому поводу, но такая реакция меня огорчила, ведь мне казалось, что этот вопрос на самом деле был пустяковым: он не давал представления о моем мышлении. В работе наличие ресурсов все же влияет на возможность справляться с задачами, в которых вы не разбираетесь. Это не история про способность зайти в кабинет со всеми необходимыми знаниями за плечами.

Что делать, если на ваш ответ отреагировали плохо?

Вы можете расстроиться из-за того, что не смогли ответить на вопрос, но не давайте эмоциям завалить все интервью. Ваша естественная реакция в такой ситуации — начать думать, как исправить оплошность. Но на самом деле вам нужно просто двигаться дальше и сосредоточиться на новых вопросах.

Думаю, если вы сталкиваетесь с подобными вопросами, вам тоже стоит узнать больше об организации. Вас будут о многом спрашивать, но вы должны оценивать эти вопросы, чтобы понять, действительно ли вы хотите работать в этой компании. Вряд ли мне подойдет работодатель, который считает, будто дата-сайентисту чрезвычайно важно уметь трижды вывести функцию генерации моментов, чтобы получить эксцесс.

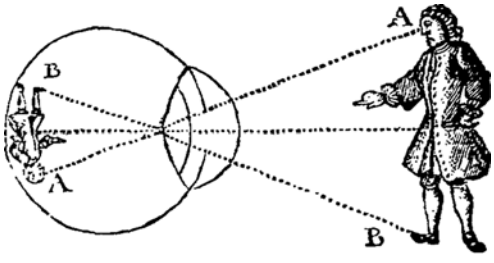
Как проведение интервью повлияло на ваше понимание положения соискателей?

Я стал гораздо внимательнее относиться к типам вопросов, которые мне задают во время интервью, а также к тем, которые задаю я сам. Прежде чем я начал проводить собеседования, я принимал все вопросы за чистую монету. Интервью было похоже на прохождение теста: меня спрашивали, а я думал, что обязательно должен ответить правильно. Я не оценивал типы задаваемых вопросов, но теперь разбираюсь в этом гораздо лучше. Интервьюер много интересуется тривиальными вещами? Хочет, чтобы я решал на доске кучу задач по программированию? А его вообще заботит, что я нашел в этой должности?

Итоги

- Процесс интервью схож в большинстве компаний, занимающихся Data Science.
- Во время интервью в офисе будьте готовы к техническим и поведенческим вопросам.
- Приготовьтесь решить кейс.
- Найдите время, чтобы попрактиковаться отвечать на распространенные вопросы.
- Определите, что вы хотите узнать о компании и должности во время интервью.

8



Оффер: знайте, на что соглашаться

В этой главе

- Рассмотрение первоначального оффера.
- Эффективные переговоры по условиям оффера.
- Выбор между двумя «хорошими» вариантами.

Поздравляю! Вы получили оффер на должность дата-сайентиста. Это большое достижение, и вам определенно стоит уделить немного времени, чтобы посмаковать момент. Вы проделали большую работу за последние месяцы или даже годы, чтобы прийти к этому.

Эта глава поможет вам ответить на полученные предложения и принять правильное решение. Даже если вы действительно в восторге, *не стоит* сразу говорить работодателю: «Да! Когда я могу приступить к работе?» Не все должности в Data Science одинаково хороши. Вы были избирательны, откликаясь на вакансии, но, возможно, во время интервью у вас появились вопросы, посеявшие сомнения в душе. Или у вас есть определенные жесткие требования, например предоставление компанией хорошей медицинской страховки для вашей семьи, поэтому вам нужно ознакомиться с деталями оффера. Даже если вы уверены, что хотите его принять, все равно не следует сразу соглашаться: нужно вести переговоры! Наступило время вашей власти, когда вы уже получили оффер, но еще не приняли его. Теперь, когда этот работодатель наконец нашел человека, который ему нравится (то есть вас!), компания должна закрыть сделку. Услуги по подбору персонала стоят очень дорого; рекрутерам и команде Data Science приходится

долго оценивать кандидатов на интервью, и за все это время организация не получает никакой выгоды при отсутствии этого (гипотетического) работника. Используйте эту возможность, чтобы спросить, что для вас важно, будь то более высокая зарплата, возможность работать раз в неделю из дома или увеличенный бюджет компании на ваше участие в конференциях.

8.1. Процесс

В целом процесс оффера выглядит примерно так:

1. *Компания сообщает вам о предстоящем оффере*. Ей нужно рассказать о предложении как можно скорее, пока вы не приняли его от другой компании.
2. *Компания делает оффер*. По электронной почте или во время звонка (а затем и электронным письмом) компания сообщает вам подробную информацию о зарплате, дате выхода на работу и других вещах, которые вы должны знать, чтобы принять решение. Обычно вам также сообщают дату, до которой вы должны определиться.
3. *Вы даете первоначальный ответ*. В разделе 8.2 мы советуем не моментально соглашаться, а выразить свою заинтересованность и попросить от нескольких дней до недели на раздумья, если вы сомневаетесь, что предложение того стоит. При следующем разговоре с работодателем вы начнете процесс переговоров (раздел 8.3).
4. *Постарайтесь выбить лучшее, что может предложить компания*. Вам могут ответить сразу же после переговоров, но часто компании требуется некоторое время, чтобы обдумать ваши требования и подтвердить, сможет ли она их удовлетворить.
5. *Решите, подходит ли вам оффер, и дайте свой окончательный ответ*.

8.2. Получение оффера

Звонок или электронное письмо с предложением обычно исходит от менеджера по подбору персонала, рекрутера или HR-а, с которым вы общались. Однако с кем бы вы ни взаимодействовали в этот момент, ваш ответ должен быть одинаковым.

Для начала скажите, как вы рады и счастливы, что получили этот оффер. Если вы будете говорить без энтузиазма, то работодатель начнет беспокоиться, что даже если вы и примете предложение, то не задержитесь надолго и не внесете весомый вклад.

Компания должна сообщить вам, что направит подробную информацию по имейлу; если этого не произошло, вам следует ее запросить. Этот нужно по двум причинам:

- У вас будет время спокойно прочитать и изучить весь пакет документов без попыток отчаянно записать все во время телефонного звонка и расшифровать свой почерк чуть позже.
- Никогда не считайте предложение официальным, пока оно не будет оформлено в письменном виде. В большинстве случаев проблем не возникнет, но вам ведь не нужно недопонимание, которое может возникнуть при звонке, например когда вы будете думать, что приняли одни условия по зарплате и льготам, а на деле окажется, что это не так.

> **ЕСЛИ ВЫ ОЧЕНЬ СИЛЬНО РАЗОЧАРОВАНЫ** Обычно мы советуем подождать, пока вы не получите всю подробную информацию о сотрудничестве в письменном виде и не потратите несколько дней на размышление, но если вы очень разочарованы предлагаемыми условиями, то, возможно, вам следует начать обсуждение. Допустим, зарплата, которую вам предложили, на 25 % меньше, чем вы рассчитывали. Вы можете начать с таких слов: «Большое спасибо за предложение. Я очень рад возможности сотрудничества и тем обязанностям, которые мне предстоит выполнять в компании Z. Но хочу быть с вами честным: зарплата, которую вы предлагаете, намного ниже, чем я ожидал. Я знаю, что в Нью-Йорке рыночная ставка для такого кандидата, как я, со степенью магистра и пятилетним опытом, находится в диапазоне от X до Y. Что мы можем сделать, чтобы моя зарплата была ближе к этому диапазону?» Чтобы обосновать свой запрос о более высокой ставке, также можете упомянуть о том, что вы получили оффер от компании-конкурента или вам предложили более высокую зарплату. Используйте эту тактику в тех случаях, когда вы готовы сразу же отказаться от вакансии из-за денег. Не применяйте ее, если вы готовы принять предложение, но хотите обсудить повышение зарплаты на 5 %. Получить на 20 % больше бывает невозможно независимо от вашей ценности и умения вести переговоры, и чем раньше вы об этом узнаете, тем лучше.

Когда вы получаете оффер, в нем должны быть указаны название должности, зарплата, любые предлагаемые опционы или акции, а также соцпакет. Если в предложении нет необходимой вам информации, например о медицинской страховке, вы также можете попросить ее добавить.

Наконец, вам должны назвать определенный промежуток времени, в течение которого нужно принять оффер. Этот срок должен составлять не менее недели; если он меньше, попросите его увеличить. Лучше всего уверенно сказать: «Мне нужно несколько дней, чтобы все обдумать». Если вы сомневаетесь, то это прекрасное время, чтобы обсудить ситуацию с кем-то еще, например с вашим партнером, членом семьи или золотой рыбкой, которая приносит удачу. Это поможет вам взглянуть на ситуацию со стороны и лучше ее оценить. Кроме того, рекрутер или менеджер сразу поймет, что надавить на вас не получится.

Иногда компании делают вам «горячее» предложение, то есть вы должны ответить менее чем за неделю; в противном случае оффер будет аннулирован. Конечно, отвечать сразу во время звонка не придется, но на раздумья вам могут дать всего сутки. Обычно вы можете перенести этот срок, сказав что-то вроде: «Я знаю, что обе стороны хотят выиграть от этой сделки. Для меня выбор следующего места работы — важное решение, и мне нужна неделя, чтобы тщательно его обдумать». В худшем случае компания по какой-то причине откажется ждать, пока вы полностью оцените оффер. Если вы оказались в такой ситуации, это очень тревожный сигнал. Работодатель, не идущий на встречу в этом случае, не будет уважать и другие ваши потребности. Выделяя вам небольшой промежуток времени, в компании знают, что вы вынуждены действовать быстро, а это вызывает тревогу и может привести к принятию неверного решения. На рынке труда есть много кандидатов, которые претендуют на работу в Data Science, а бизнесу все еще сложно найти подходящих сотрудников, поэтому если с вами обращаются подобным образом, в компании, скорее всего, происходит что-то неладное.

Если вы проходили интервью в других организациях, сообщите им, что вы получили оффер, а также озвучьте дату, когда вам нужно дать ответ. Если вы находитесь на последнем этапе собеседования, то эти компании могут ускорить процесс, чтобы пригласить вас до того, как вам придется давать ответ по текущему предложению. Сообщать об этом другим организациям совершенно нормально; они это оценят. Когда вы будете снова с ними общаться, еще раз подчеркните, насколько вам нравится компания и та работа, которую вы там будете делать. Некоторые работодатели могут никак не отреагировать. Что ж, по крайней мере вы дадите им шанс. Опять же, так как хорошего дата-сайентиста найти непросто, если компания находит того, кто ей нравится, она обычно действует быстро, не затягивая процесс.

8.3. Переговоры

Многие ненавидят обсуждать офферы. Частично это потому, что обсуждение воспринимается как игра с нулевой суммой: если выиграете вы, проиграет компания. Другая причина в том, что такие переговоры заставляют чувствовать себя эгоистичным и жадным, особенно если этот оффер лучше, чем ваша нынешняя работа. В такой ситуации легко решить, будто вам повезло, что вы хоть кому-то подошли, и не осознать, что вы заслуживаете больше предложенного. Но, если вы находитесь в такой ситуации, то обязаны сделать все возможное, чтобы получить от компании максимум.

Вы — да, именно вы тот самый человек, который лучше всего подходит для этой работы. Все мы сталкиваемся с синдромом самозванца, но работодатель видит в вас того, кто ему нужен на эту должность. Как мы уже говорили в начале главы, сейчас самое удачное время для ведения переговоров. И компании ждут их! Вы должны быть готовы. Зарплата не связана с вашими прежними ставками или с тем, что

компания может изначально предложить; напротив, она зависит от того, что рынок предлагает людям с вашим набором навыков. Убедитесь, что вам платят столько же, сколько вашим коллегам, что общая сумма вознаграждения соответствует вашим ожиданиям и что вы получаете от фирмы наиболее важные для вас льготы. У вас есть шанс договориться об увеличении зарплаты на 5 % за один пятиминутный телефонный звонок; зато вы надолго обеспечите себе комфортную жизнь.

В сфере Data Science вопрос о необходимости вести переговоры стоит особенно остро из-за огромной разницы в зарплатах. Поскольку это новая область, в которой много разных направлений, какие-либо стандарты обязанностей сотрудников отсутствуют. Два человека с одинаковым набором навыков могут получать совершенно разные деньги, при этом один называет себя специалистом по данным, а другой — инженером по машинному обучению. Эти различия в зарплате усугубляются по мере того, как люди переходят на новую высокооплачиваемую работу. На данный момент соотношение между суммой, которую человек зарабатывает, и его квалификацией и способностями весьма незначительно.

8.3.1. Что можно обсуждать?

Первое, о чем многие думают при обсуждении оффера, — это зарплата. Прежде чем вы дойдете до последних этапов интервью, вам следует узнать размеры зарплаты не просто дата-сайентистов в целом: постарайтесь выяснить, сколько получают специалисты по данным именно в этой отрасли, в конкретном городе и конкретной компании, если они в ней уже есть. Несмотря на то что в оффере вам могут предложить значительную сумму по сравнению с прошлым местом работы, помните, что вам должны платить так же, как вашим коллегам в новой компании. Вы не обязаны сообщать компании о вашей текущей зарплате во время интервью, а в некоторых городах и штатах это даже незаконно. Если вы зайдете на Glassdoor и обнаружите, что средний дата-сайентист зарабатывает вдвое больше вас, это прекрасно.

Если на собеседовании вас спросят, сколько вы сейчас зарабатываете или какие у вас зарплатные ожидания, постарайтесь уйти от ответа, поскольку иначе вы рискуете получить более низкое предложение, так как работодатель будет знать, что оправдает ваши ожидания. Вы можете попробовать ответить: «Сейчас я больше сосредоточен на поиске должности, которая будет хорошо соответствовать моим навыкам и опыту. Я уверен, что если мы подойдем друг другу, то сможем договориться и о соцпакете». Если человек продолжает давить на вас и говорит, что хочет убедиться в отсутствии у вас ожиданий чего-то выходящего за рамки предлагаемого диапазона, то можете ответить так: «Я понимаю, что вы хотите, чтобы мои ожидания соответствовали вашим возможностям. Уточните, пожалуйста, каков диапазон для этой должности?» Вы также можете сказать: «Мне хотелось бы задать еще несколько вопросов [менеджеру по персоналу/старшему

дата-сайентисту/любому человеку, с которым вы непосредственно сейчас не разговариваете], чтобы лучше понять, что предполагает эта должность, прежде чем смогу озвучить реалистичные ожидания». Если человек отказывается назвать диапазон и не идет вам навстречу, пока вы не согласитесь назвать сумму, это плохой знак. Если вам очень понравилась компания, а от ответа уйти не получилось, скажите что-нибудь вроде этого: «Что касается этого вопроса, то у меня довольнь гибкие ожидания в зависимости от общего соцпакета и я знаю, что от $X до $Y — стандартная сумма для человека с моими опытом и образованием».

Вы не должны стыдиться просьб о разумной оплате. То, что вы считаете обоснованным, не должно определяться вашим прошлым заработком, а только тем, сколько вы планируете получать на новой должности. Компании часто предлагают меньше, ожидая, что в процессе обсуждения вы договоритесь о той сумме, которую они рассчитывают платить новому сотруднику, поэтому повышение как минимум на 5 % одобряют достаточно часто.

Но договориться можно не только о зарплате. Во-первых, есть и другие прямые денежные выплаты, например поощрительная премия при устройстве на работу, пакет релокации и акции. Работодателю проще выплатить премию новому сотруднику, чем повысить ему зарплату, так как премия — это разовый бонус. То же самое и с расходами на релокацию; если вы собираетесь переехать, спросите о политике компании в отношении переезда. Если организация большая, скорее всего, у нее есть стандартный релокационный пакет, а порой и контракт с определенным перевозчиком. Даже небольшие фирмы могут что-то вам предложить; не факт, что вы узнаете об этих бонусах, если не спросите сами.

Но вы можете пойти еще дальше. Вернитесь к тому, о чем думали, находясь в поиске работы. Что для вас важно? О некоторых вещах трудно договориться, например варианты медицинской страховки и пенсионного плана часто одинаковы для всех, кто занимает определенную должность. Вы должны учитывать это на ранних этапах. Отзывы сотрудников о пакете льгот в устоявшихся компаниях обычно можно найти на сайте Glassdoor. Но есть много других вещей, о которых вы можете попросить, например:

- гибкий или удаленный график работы;
- более ранняя оценка вашей деятельности (через полгода, а не год), которая поможет быстрее получить повышение;
- образование за счет компании;
- бюджет на участие в конференциях.

При этом помните, что у компании тоже есть ограничения. Например, у некоммерческих организаций вряд ли будет большая вилка зарплат, зато они могут быть гибкими в отношении рабочего времени или отпуска. Компания с распределенной по отделам командой Data Science с большей вероятностью позволит

вам работать из дома несколько дней в месяц, чем та, где все работают в одном офисе. Убедитесь в том, что ваши просьбы зафиксированы на бумаге и приняты. Вы ведь не хотите о чем-то договориться и не получить, принимая работу. Также имейте в виду, что подобные не выраженные в деньгах возможности являются лишь дополнением. Если зарплата намного ниже ваших ожиданий, эти изменения вряд ли сделают предложение приемлемым.

Кейтлин Хадон (Caitlin Hudon) о синдроме самозванца

При серьезных изменениях в карьере многие люди — даже опытные ведущие дата-сайентисты! — испытывают приступы синдрома самозванца. *Синдром самозванца* — это чувство сомнения в своих достижениях и беспокойство, что вас разоблачат как мошенника. В частности, переговоры о зарплате — дело непростое: это очень важный вопрос, упражнение в определении вашей ценности в буквальном смысле. Ваше мнение о себе почти наверняка будет влиять на принятие решения, поэтому важно побороть синдром самозванца, чтобы быть объективным и настойчивым при демонстрации своих способностей.

Синдром самозванца особенно распространен среди дата-сайентистов. Разные люди скажут вам, что специалист по данным — это некая комбинация аналитика/статистика/инженера/специалиста МО/ визуализатора/специалиста БД/бизнес-эксперта, при том что в действительности каждая из этих ролей является самостоятельной. Кроме того, в этой постоянно развивающейся сфере необходимость оставаться в курсе новейших технологий может накладывать дополнительный стресс. Итак, в двух словах: у нас есть люди с разным опытом, приходящие в новую область, границы которой четко не обозначены (что приводит к неизбежным пробелам в знаниях об этой области в целом) и где технологии меняются со скоростью света. Если вам кажется, что одному человеку будет сложно с этим справиться, значит, так оно и есть!

Таким образом, мы подобрались к моему подходу в борьбе с синдромом самозванца, суть которого — сосредоточиться на своем уникальном опыте, делающем меня особенной, а не сравнивать себя с недостижимым идеалом дата-сайентиста. Я приняла для себя идею, что никогда не смогу выучить все, что нужно знать в этой сфере, не узнаю все алгоритмы, технологии, крутые пакеты или языки программирования — и это нормально. (Самое замечательное в такой разнообразной, развивающейся сфере то, что никто другой тоже не сможет, и это нормально!)

Более того, мы с вами знаем и умеем такое, чего не знают и не умеют другие. Ваши навыки и опыт делают вас похожими на остальных специалистов и в то же время выделяют вас среди остальных. Сосредоточьтесь на собственном уникальном опыте. В конце концов, вы его получили упорным трудом, и важно помнить, что он отличает вас от других специалистов по данным.

Когда вы говорите о зарплате и сталкиваетесь с синдромом самозванца, подумайте обо всех навыках, которые вы приобрели, о задачах, которые решили; подумайте о ценном вкладе, который вы внесете в свою будущую команду. Ваш опыт ценен, и платить за него нужно справедливо. Не бойтесь просить о том, чего заслуживаете.

8.3.2. О какой сумме договариваться

Ваш лучший рычаг на переговорах — это конкурирующее предложение. Оно дает понять компании, что рынок готов больше платить за ваши услуги и вам есть куда пойти. Если у вас окажется два оффера, лучше всего сказать так: «Я бы предпочел работать с вами, но компания ABC предлагает мне гораздо больше. Вы можете предложить столько же?» Не лгите: если вы попытаетесь сказать это обеим компаниям, это может легко обернуться не в вашу пользу.

Еще один рычаг — ваша текущая работа. Большинство компаний понимает, что вы не захотите получать меньше. Если вы относительно довольны или ваша фирма дает больше преимуществ, воспользуйтесь этим. Даже если вам предложили более высокую зарплату, не забудьте учесть преимущества текущего соцпакета. Допустим, нынешний работодатель предлагает 3 %-ный пенсионный план, а новый — ничего. Получается, что зарплата на текущем месте работы на 3 % выше. Или, возможно, ваша нынешняя компания полностью оплачивает медицинскую страховку, тогда как в новой вам придется платить по $200 в месяц за страхование вашей семьи. Обязательно учтите это, особенно если переходите из более низкооплачиваемой области: хотя ваш прямой доход существенно вырастет, вы потеряете льготы, которые получали прежде. Можно указать это в качестве причин, по которым вы просите более высокую зарплату.

Оценивая свои навыки в денежном эквиваленте, подумайте, насколько уникален ваш опыт для нового работодателя. Вам не обязательно быть опытным специалистом. Конечно, круто, если вы один из трех выпускников Стэнфорда, получивших в этом году докторскую степень в области искусственного интеллекта. Но предположим, что раньше вы работали в продажах, а теперь собираетесь стать дата-сайентистом, который помогает повысить уровень этих самых продаж. Знание этой сферы — огромное преимущество, которого может не быть даже у более опытных специалистов по данным. Чем больше вы уверены, что эта вакансия создана именно для вас, тем больше рычагов влияния сможете применить.

Иногда бывает трудно представить полную картину оффера. Например, если вам придется сменить место жительства, то следует подумать о тратах на переезд, а также об изменении размера повседневных расходов. Зарплата в $90 000 в Хьюстоне даст намного больше, чем $95 000 в Нью-Йорке. Сложите все преимущества. Хорошая медицинская страховка или пенсионный план может быть лучше, чем тысячи или даже десятки тысяч долларов. В общем, не упускайте из виду лес за деревом зарплат.

Если вы провели переговоры и получили все, о чем просили, компания ждет, что вы примите оффер! Начинайте обсуждение только в том случае, если собираетесь принять предложение при соблюдении всех ваших условий. Иначе зачем тогда все это затевать? Вообще-то, на это можно пойти, если у вас уже есть оффер от другой компании, который вы очень хотите принять, и чем лучше предложение

ее конкурента, тем больше вы сможете обсуждать условия. Однако это чрезвычайно рискованная тактика: будьте осторожны и не сожгите ненароком мосты. Даже если вы никогда не пойдете работать в компанию, которую использовали в качестве рычага давления, интервьюеры или сотрудники, сделавшие оффер, могут вас запомнить.

Компании очень и очень редко отзывают оффер из-за того, что вы решились на переговоры. Если это произошло, то работать у них точно не стоит. Совершенно нормально уважительно обговаривать условия, а вот отклонение кандидата по этой причине — очень тревожный звоночек. Считайте, что вам удалось избежать неприятностей.

Краткое руководство по ограниченным акциям компании, опционам и планам покупки акций для сотрудников

Эта заметка намеренно названа «кратким руководством», а не «всесторонним обзором». Мы настоятельно рекомендуем дополнительно изучить всю информацию в вашем оффере.

Как правило, все перечисленное ниже распространяется на четыре года вперед с ограничением в один год. То есть если вы уволитесь раньше, чем через год, то ничего не получите. Зато сразу по истечении этого срока вы получите всю накопившуюся сумму. Затем бонусы, как правило, будут переходить каждый квартал или месяц в течение следующих трех лет:

- *Акции с ограничениями* (*ACO*). В оффере укажут стоимость передаваемых вам ACO в долларах. Если вы разделите указанную стоимость на цену акции в момент выдачи оффера, то получите количество предлагаемых акций. Если они вырастут в цене, повысится и компенсация. Как правило, акции передаются пакетом; обычно компания удерживает часть налоговых отчислений в виде акций.

  Предположим, в предложении значится сумма в $40 000, переходящая вам в течение четырех лет с годовым порогом. В настоящий момент акции торгуются по $100 за штуку, соответственно, вы получите 100 акций через год и далее по 25 акций каждый квартал. После того как вы проработаете в компании год, вы получите 65 акций (35 будет удержано в виде налогов). После этого вы можете делать с акциями все, что считаете нужным, пока соблюдаете правила компании (например, продавать акции обычно можно только в определенные периоды). Если вы работаете в частной компании, передача акций происходит по той же схеме, но обычно продать свои акции нельзя, пока компания не станет публичной или не будет кем-то выкуплена.
- *Опционы на акции*. Опционы дают вам возможность купить некоторое количество акций по определенной цене реализации, которая обычно равна справедливой рыночной стоимости акций на момент предоставления опциона. Если позже цена акции вырастет, это будет здорово: когда можно купить акции по $10 за штуку, а они торгуются по $30, вы мгновенно заработаете $20, просто купив и продав акции сразу же! Однако если акция не торгуется выше цены реализации опциона,

она ничего не стоит: вы не станете использовать опцион на покупку акций по $10, если есть возможность купить их на рынке по 5.

Опционы особенно выгодны для старых сотрудников выросших компаний, но они также могут быть похожи на лотерейные билеты. Пока ваша компания остается частной, стоимость опционов ограничена, и даже если компания станет публичной, акции могут в конечном итоге торговаться ниже стоимости вашего опциона. В отличие от ACO, которые всегда будут иметь определенную ценность, опционы могут никогда не принести вам денег.

- *Программы предоставления акций работникам компании* (*employee stock purchase plans, ESOP*). Такие программы позволяют покупать акции компании со скидкой. Накопление средств для покупки происходит за счет удержания из заработной платы. По накоплении нужной суммы наступает дата реализации, и ваши деньги используются для покупки акций. У этих программ есть два преимущества:
- Скидка на стоимость акций, которая может достигать 15 %.
- Ретроспективная цена реализации. Если с начала срока действия предложения до даты покупки цена выросла, вы купите акции дешевле.

Предположим, что на начало срока действия предложения акция компании стоила $10. Вы отчислили $9000 в течение года и достигли даты реализации. При размере скидки в 10 % вы можете купить 1000 акций ($9000/$9). Если сейчас цена акции составляет $20, то 1000 акций будут стоить $20 000, то есть при продаже можно получить $11 000 прибыли!

8.4. Тактика переговоров

Теперь, когда мы рассмотрели общую картину, перейдем к некоторым практическим советам по ведению переговоров.

- *Не забудьте для начала выразить благодарность и радость*. Надеемся, что это так! Работодатель должен почувствовать, что вы на его стороне и готовы работать вместе. Если вы не дадите этого понять, к вам будут относиться настороженно.
- *Подготовьтесь*. Перед звонком запишите, что именно вы хотите изменить в оффере и на какой доход готовы согласиться. В спешке можно легко выпалить число меньше желаемого, чтобы порадовать человека на другом конце провода. С заметками под рукой этого проще избежать.
- *Слушайте, что вам говорит собеседник*. Если он подчеркивает, что зарплата не обсуждается, но этот вопрос для вас очень важен, спросите о дополнительных бонусах или поощрительной премии для новых сотрудников. Старайтесь найти решение совместными усилиями, проявляйте инициативу. Переговоры — это не обязательно игра с нулевой суммой, хотя порой кажется именно так! Что-то очень важное для вас может быть непринципиально для работодателя, и тогда ему легко пойти вам на встречу. Помните: компания хочет, чтобы вы стремились к успеху и были счастливы на своей должности.

- *Не производите впечатление человека, которого интересуют только деньги* (даже если это и так). Это может плохо закончиться. Вы должны показать, что вас мотивируют будущие задачи, миссия компании и новые коллеги. Если рекрутер решит, что вы пришли только ради денег, то насторожится и начнет беспокоиться, что вы ухватитесь за возможность уйти ради более высокой зарплаты.
- *Старайтесь сосредоточиться на коллективных целях*. Вместо фразы: «Мне действительно нужно больше опционов» скажите: «Я очень рад перспективам долгосрочного роста в компании и возможности помочь вам добиться успеха. Дополнительные опционы сделают меня еще более заинтересованным в успехе и помогут работать с большей отдачей».
- *Объединяйте свои желания, а не раскладывайте по пунктам*. Так у рекрутера будет полная картина ваших предпочтений и не возникнет ощущения, что как только вы решите одну проблему, тут же появится другая. При этом можно не перечислять весь список сразу. Например, если вы просите зарплату повыше, опционов побольше и в придачу возможность работать из дома один день в неделю, а компания не согласна с пунктом про зарплату, то взамен можно попросить премию для новых сотрудников. По возможности также укажите, насколько важен для вас каждый пункт.
- *Избегайте самоуничижения*. Прочли в разделе чуть выше, насколько вы хороши? Прочтите еще раз. Не обижайте себя фразами вроде: «Я знаю, что раньше не был дата-сайентистом» или «Я знаю, что у меня нет докторской степени, но…» У вас есть именно то, что ищет работодатель, иначе он не прислал бы вам оффер!

В приложениях мы перечислили несколько наших любимых книг и статей, где можно найти гораздо больше советов и исследований на тему переговоров.

> **ПЕРЕГОВОРЫ ДЛЯ ЖЕНЩИН** Согласно распространенной теории, «женщины не задают вопросов». Эта теория объясняла одну из причин, по которой мужчинам платят больше. Дело в том, что женщины не обсуждают условия оффера и не просят повышения. Однако недавнее исследование показало, что по крайней мере в некоторых областях они ведут переговоры не реже мужчин; у них просто меньше шансов на успех. К сожалению, существует некоторая предвзятость. Некоторые из тактик, которые мы предложили в этом разделе, например сосредоточиться на коллективных целях (говорить «мы» вместо «я»), особенно подойдут для женщин.

8.5. Как выбрать между двумя «хорошими» офферами

Получить два (или больше!) хороших оффера — большая проблема! Но выбор все же придется сделать. Конечно, не факт, что это будет трудно. Иногда одна работа настолько нравится, что принять решение легко. Но что делать, если это не так?

Первый шаг — вернуться к разделу 8.3. Обсуждайте условия! Если действительно одна компания вам нравится больше, чем другая, но есть какое-то препятствие, посмотрите, может ли работодатель что-то с этим сделать. Говорите честно о своих потребностях. Как мы уже говорили, конкурирующий оффер является отличным рычагом воздействия и повышает вероятность получить желаемое.

Можно попросить о дополнительной встрече с менеджером, проводящим набор, или с потенциальными коллегами, чтобы получить больше информации. Возможно, вы не успели поинтересоваться средним количеством совещаний в неделю, числом команд, с которыми придется работать, способом расставления приоритетов или чем-то еще, что помогло бы вам оценить степень привлекательности компании.

Когда вы рассматриваете оффер, подумайте о будущем. Бывают и вопросы, которые необходимо решать в ближайшее время: если на вас висит большой кредит за обучение или вы планируете создать семью, возможно, следует принять оффер с наибольшей зарплатой, чтобы начать копить или выплачивать долг. Но если вам повезло и подобных вопросов решать не нужно, лучше максимально сосредоточьтесь на долгосрочном потенциале. Чем вам предстоит заниматься в каждой компании? Если один работодатель оплатит ваше обучение, благодаря которому можно продвинуться на два шага по карьерной лестнице, а другой заставит вас делать то, чему вы научились в первый день на курсах, не следует хвататься за оффер последнего лишь потому, что он предлагает более выгодный отпускной пакет. Помните, что зарплаты в Data Science могут варьировать на целых 30–40 % в зависимости от должности. Если у вас скромное резюме и эта работа — отличный шанс проявить себя, можно согласиться на более низкую зарплату и поработать год или два.

Наконец, смело рассматривайте менее значимые факторы. Может быть, у одной из компаний более просторный офис или добираться до него гораздо удобнее? Если обе организации соответствуют вашим минимальным критериям и похожи в плане наиболее важных для вас факторов, подумайте о деталях попроще.

В подобных ситуациях рассматривайте несколько вариантов. Вы беретесь за увлекательную, но рискованную работу в стартапе или отправляетесь трудиться на надежного государственного подрядчика? Принимаете ли вы оффер, по условиям которого вам придется стать руководителем, или продолжаете работать в качестве независимого специалиста?

Часто объективно выбрать лучший путь невозможно. Вы не поймете, насколько вам подходит новая работа, пока не примете оффер. Не узнаете, понравится ли вам руководить людьми, пока не попробуете себя в этой роли. Такова правда жизни.

Если вам трудно принять решение, просто помните, что результат зависит от качества информации у вас на руках. Вы не можете смотреть в будущее. Если вы приняли решение и в итоге оказались недовольны, ищите другой выход. Можно уволиться, вернуться в родной город или снова стать независимым специали-

стом. Жизнь неоднозначна, и можно извлечь много полезных уроков даже из тех ситуаций, которые ни к чему не привели.

Когда вы все же примете решение, отказывайтесь от любых других предложений изящно. Это будет вежливо с вашей стороны, да и мир Data Science тесен. Не стоит отвергать оффер со словами: «Не могу поверить, что у вас такое невыгодное предложение; ужасная и совершенно неэтичная компания». А то вдруг через пять лет вы узнаете, что человек, на которого вы так громко кричали, набирает персонал в компанию вашей мечты.

8.6. Интервью с Брук Уотсон Мадубуонву, старшим дата-сайентистом в ACLU

Брук Уотсон Мадубуонву (Brooke Watson Madubuonwu) работает старшим специалистом по данным в Американском союзе защиты гражданских свобод (ACLU) и занимается количественным анализом для судебных и правозащитных групп по вопросам, связанным с гражданскими правами. У нее есть степень магистра эпидемиологии и опыт работы в роли инженера-исследователя.

Что кроме зарплаты необходимо учитывать при рассмотрении оффера?

Для начала вы должны расставить приоритеты. Сравнить зарплату легко, но нужно также оценивать, как будет работаться в компании каждый день или месяц. Я предпочитаю рассматривать три категории: образ жизни, обучение и ценности.

Образ жизни — это то, как ваша работа вписывается в другие ваши планы. Где я буду жить? Могу ли трудиться удаленно? А что насчет работы по вечерам и в выходные? Смогу ли я путешествовать? Придется ли ездить в командировки? Какие варианты медицинской страховки, отпуска по уходу за детьми и пенсионные программы мне доступны? Смогу ли я совмещать работу и одновременно заботиться о семье? Это первое.

Второе — это обучение. Буду ли я развиваться на этой должности? Какие системы мне в этом помогут? Есть ли чему учиться у начальника и команды?

Наконец, я думаю о ценностях и миссии организации и конкретной команды. Работают ли они над тем, что соответствует моим ценностям? Поощряет ли команда инклюзивность? Что у меня в приоритете: продукт или команда? Каждый из этих трех аспектов может быть по-разному важен в зависимости от времени.

Я говорила с руководителем об изменении названия своей должности, а вы можете обсуждать такие вещи, как обучение и участие в конференциях, возможность работать несколько дней дома, командировки или акционерные доли. Data Science — это обширная сфера, которая постоянно развивается, поэтому я думаю, что для улучшения своих навыков важно экономить время в течение рабочего дня.

Как вы готовитесь к переговорам?

Раньше мне было очень неловко говорить о своей ценности и продвигать ее. Я пыталась отучиться от этого на протяжении всей своей карьеры. Что действительно может быть полезно, так это наличие союзников. Потренируйтесь вести переговоры с близким другом, попросите его выступать в роли вас. Многим людям, в том числе и мне, гораздо легче защищать друзей, чем самих себя. Описание кем-то ваших сильных сторон и потребностей может неплохо мотивировать, и это действительно помогает. Спросите себя: «Что я могу рассказать о человеке, которого люблю, если знаю, что он отлично подходит для этой работы, и хочу, чтобы у него все получилось?» Вот как вам следует говорить о себе.

Что делать, если у вас есть один оффер, но вы все еще ждете другого?

Когда вы прошли все этапы интервью или их бóльшую часть, то можете упомянуть, что у вас есть еще один оффер. Также можно выиграть время, если попросить у компании, предложившей сотрудничество, неделю или две, чтобы изучить детали и обсудить предложение с семьей. Дальнейшие переговоры, подробные вопросы о соцпакете и просьба поговорить с членами команды о работе или культуре компании — все это можно пускать в ход ради двойной цели: собрать побольше информации о должности и выиграть время.

Но если вам предстоит еще несколько раундов интервью, возможно, ускорить процесс не получится и придется сравнивать полученный оффер с текущей ситуацией. Если вы только что закончили учебу или сидите без работы, выбирать особо не приходится. Но если вы трудоустроены и этот оффер вам не интересен, пожалуй, лучше будет остаться у нынешнего работодателя до тех пор, пока не найдется более подходяще место. Первое предложение, которое я получила в области науки о данных, мне не подошло, и, несмотря на то, что мне очень хотелось перейти в эту сферу, пришлось еще немного подождать.

Ваш последний совет начинающим дата-сайентистам и джунам?

Я бы посоветовала относиться к названиям должностей без излишней предвзятости. Думаю, что «дата-сайентист» звучит привлекательно, особенно для тех, кто специализируется в этой области; когда я училась в колледже, такой специализации еще не было. Некоторые считают, что если они не получат работу специалиста по данным сразу после окончания учебы, то это провал. Но повседневные обязанности дата-сайентиста могут быть у аналитиков, исследователей и многих других работников; эти роли могут стать для вас действительно отличным стартом для оттачивания своих навыков. Я училась программированию во время работы лаборантом, а затем и инженером-исследователем за годы до того, как меня стали называть дата-сайентистом, при этом мои обязанности были не менее интересными. На старших курсах я работала специалистом по вводу данных, и даже это

помогло мне сформировать представление о том, как решения о сборе данных влияют на аналитические возможности. Нет плохих работ, если вы готовы учиться.

Итоги

- Не принимайте оффер сразу: получите подробную информацию в письменной форме и попросите время, чтобы ознакомиться с деталями.
- Обсуждайте, обсуждайте и еще раз обсуждайте! Можно попросить более высокую зарплату или другие финансовые льготы, но не стоит забывать о таких вещах, как гибкий график работы и бюджет на участие в конференциях.
- Взвешивая два хороших оффера, рассматривайте их в долгосрочной перспективе; не думайте только о стартовой зарплате.

Материалы к главам 5–8

Книги

I Will Teach You to be Rich, 2nd ed., Ramit Sethi (Workman Publishing Company)

Эта франшиза представляет собой блог, книгу, видео и несколько других медиа, и все они пригодятся при поиске работы. Рамит Сетхи (Ramit Sethi) дает потрясающие советы о прохождении интервью и способах давать хорошие ответы. Здесь вы также найдете информацию о том, как обсуждать зарплату, просить повышение и общаться на другие непростые темы, которые всплывают, когда нужно отстаивать свою позицию.

What Color Is Your Parachute? 2020: A Practical Manual for Job-Hunters and Career-Changers, Richard N. Bolles (Ten Speed Press)

Эта книга пригодится тем, кто ищет новую работу. Как и в главах 5–8, в ней раскрываются темы определения требований к работе, поиска вакансий, составления резюме, прохождения интервью и ведения переговоров. Из книги можно узнать, как смотреть на поиск работы более широко, при том что в ней нет ничего специфического для технических профессий (не говоря уже о работе в Data Science).

Cracking the Coding Interview, Gayle Laakmann McDowell (CareerCup)

Эта книга призвана помочь программистам устроиться на работу. В ней есть почти 200 вопросов интервью с примерами ответов. Несмотря на то что большинство не относятся к Data Science, многие из них все же окажутся полезны, особенно если вы ищете место инженера МО, где нужно довольно хорошо программировать.

Блоги и курсы

«Advice on applying to Data Science jobs», Jason Goodman

http://mng.bz/POlv

Это отличная публикация об уроках, извлеченных из откликов на вакансии в Data Science. Она охватывает многие концепции, описанные в главах 5–8, но с личным взглядом автора на происходящее.

«How to write a cover letter: The all-time best tips», Muse Editor

http://mng.bz/Jzja

Хотя в главе 6 мы подробно рассказали, как писать сопроводительное письмо, было бы неплохо посмотреть на этот вопрос с другой стороны. В этой статье рассматриваются темы, которых нет у нас, например как перестать думать, будто вы хвастаетесь.

«Up-level your résumé», Kristen Kehrer

https://datamovesme.com/course_description/up-level-your-resume

Кристен Керер, которая отвечала на наши вопросы в конце главы 6, создала этот платный курс для помощи начинающим дата-сайентистам оптимизировать свое резюме и сопроводительное письмо, чтобы успешно пройти автоматизированную систему для подбора персонала (роботов) и рекрутеров. Если вас редко приглашают на первый этап интервью, присмотритесь к этому курсу.

«How to quantify your resume bullets (when you don't work with numbers)», Lily Zhang

https://www.themuse.com/advice/how-to-quantify-your-resume-bullets-when-you-dontwork-with-numbers

Если наш совет называть цифры в пунктах резюме вызвал у вас трудности, эта статья может вам помочь. В ней описаны три способа количественной оценки вашего опыта: диапазон, частота и масштаб.

«How women can get what they want in a negotiation», Suzanne de Janasz and Beth Cabrera

https://hbr.org/2018/08/how-women-can-get-what-they-want-in-a-negotiation

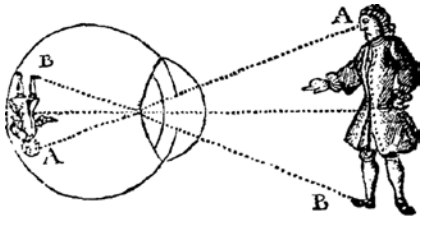
В этой публикации рассказывается, как женщинам преодолеть присущие им предубеждения при обсуждении заработной платы и получении оффера. Женщинам стоит задуматься об этой теме, потому что польза от умения эффективно вести переговоры растет на протяжении карьерного пути.

«Ten rules for negotiating a job offer», Haseeb Qureshi

https://haseebq.com/my-ten-rules-for-negotiating-a-job-offer

Эта статья объясняет, как эффективно обсуждать оффер. Если на этапе переговоров вы располагаете правильной информацией, то можете добиться разницы в тысячи долларов, поэтому будет нелишним прочесть этот пост.

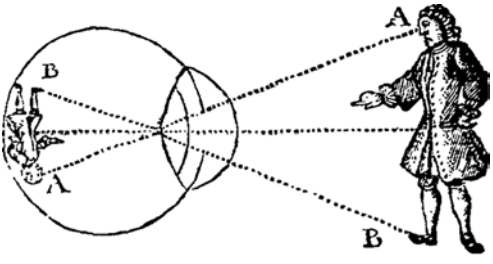
Часть 3 Осваиваемся в Data Science



Приступить к своей первой работе в Data Science — настоящее достижение, но это лишь начало пути. Трудиться на компанию в роли специалиста совсем не то же самое, что заниматься данными на курсах или в качестве хобби. Возможно, вам придется освоить множество новых концепций, от корпоративного этикета до правильного способа внедрения кода в производство. Огромная разница между ожиданиями и реальностью рабочих будней может быть ошеломляющей. Эта часть книги послужит в качестве спасательной подушки для смягчения удара. Прочитав ее, вы узнаете, чего следует ожидать от работы в Data Science, и будете лучше подготовлены.

Глава 9 посвящена стартовым месяцам: начиная с первых дней, когда порой чувствуешь абсолютную растерянность, и до осваивания на новом месте, когда вы уже лучше знаете свои обязанности и коллег. Глава 10 дает руководство по созданию хорошего анализа (основная часть обязанностей большинства дата-сайентистов) путем составления и выполнения изначального плана. В главе 11 мы расскажем об использовании моделей машинного обучения и их внедрении в производство, а также введем такие понятия, как «модульное тестирование», которые необходимы инженерам данных. Глава 12 — это глубокое погружение в чрезвычайно важную задачу по работе со стейкхолдерами, которой часто приходится заниматься дата-сайентистам и с которой возникает больше всего трудностей.

9



Первые месяцы на работе

В этой главе

- Чего ожидать в первые несколько недель работы дата-сайентистом.
- Как стать продуктивным, выстраивая отношения и задавая вопросы.
- Что делать, если вы попали в плохую рабочую среду.

В этой главе мы расскажем, чего ожидать в первые несколько месяцев и как провести их с пользой. Именно это время окажет огромное влияние на то, как будет продвигаться работа; это ваш шанс создать систему и сеть поддержки, которая приведет вас к успеху. Хотя каждая должность в Data Science индивидуальна, некоторые общие принципы применимы ко всем ролям.

В начале работы инстинктивно хочется сделать как можно больше. С этим желанием нужно бороться. Вы должны быть уверены, что не просто выполняете задачи, а делаете это правильно. Первые дни — это наиболее подходящий момент для вопросов о том, как что-то делать, потому что от вас не ждут знания процессов компании. Руководители иногда забывают, что вы можете не знать того, что знал ваш предшественник, поэтому вам могут поручить задачу, в которой вы совершенно не разбираетесь. Возможно, вы сможете обхитрить систему и справиться с чем-то самостоятельно, но гораздо лучше задать вопросы заранее и найти подход к рабочему процессу.

9.1. Первый месяц

То, как пройдет ваш первый месяц, будет зависеть от типа компании. У организаций разного масштаба подход к новым сотрудникам совершенно отличается. На рис. 9.1 показано, чего можно ожидать от двух компаний: большой фирмы с массой дата-сайентистов и маленькой, у которой нет или почти нет такой команды. (В главе 2 такими примерами были КИТк и Seg-Metra соответственно.) Эти два примера находятся на разных концах спектра, однако ваша организация, скорее всего, окажется где-то посередине.

Рис. 9.1. Онбординг в крупной организации похож на прохождение заводского конвейера, тогда как в небольших компаниях она менее организованна. (Эмодзи Twitter из проекта Twemoji)

9.1.1. Онбординг в крупной организации: хорошо отлаженный процесс

Вы один из десятков людей, приступающих на этой неделе к работе. За семь дней до этого вы получили имейл с информацией о том, куда идти, когда приехать и что нужно взять с собой. Теперь для вас начинается формальный многодневный процесс онбординга и взаимодействия с людьми из разных отделов. Вам выдадут ноутбук, который вы настроите самостоятельно. Вы прослушаете презентации о корпоративной культуре, кадровой политике и устройстве компании. Все работает как часы, до вас в организацию были приняты тысячи сотрудников.

Что касается Data Science, вам помогут настроить среды программирования. Вероятно, вам выдадут контрольный список или обширную документацию с указанием всего, что нужно сделать, чтобы получить доступ к данным. Также можно ознакомиться с центральным репозиторием старых отчетов и документации с данными. Никто не ждет от вас всего и сразу: коллеги рады, что вы присоединились к команде, и знают, что вам нужно время, чтобы освоиться. Вам потребуется несколько недель, чтобы всему обучиться и получить одобрение доступа к системам. Необходимость так долго ждать может разочаровать: вы

же так хотите почувствовать себя продуктивным! Но медленный старт в такой сфере — это нормально.

Если вам назначили одну или несколько задач, отнеситесь к ним серьезно, но сосредоточьтесь больше на процессе, а не на результате. У устоявшихся команд часто есть свои особенности, которые необходимо принять. На этом этапе не только можно, но даже нужно задавать вопросы; это залог дальнейшего успеха. Первые несколько месяцев — это шанс увидеть, что было сделано до вас, и узнать, в каком ритме работали ваши коллеги.

9.1.2. Онбординг новых сотрудников в небольшой компании: «Онбординг? Нет, не слышали»

«О, ты выходишь уже сегодня?» Если вы пришли в небольшой стартап, не удивляйтесь, если будет готово не все, включая ноутбук. Возможно, вам самостоятельно придется выяснять, как получить доступ к данным. Когда вы все-таки войдете в систему, может оказаться, что данные недостаточно оптимизированы для работы и что SQL-запрос к небольшой таблице из 100 000 строк занимает шесть минут. Инструктаж нового сотрудника для ознакомления с компанией может не проводиться несколько недель, если он вообще будет. Это может быть связано с малым числом новичков, для которых требуется это мероприятие.

Каких-либо стандартов нет. Никто не будет указывать, какой язык программирования использовать или как подходить к анализу и структурировать его. Зато результат с вас спросят довольно скоро. В отличие от работы в крупной организации, вам не придется переживать за свою продуктивность; вас попросят сразу же приступать к работе. А вот о чем вам действительно стоит задуматься, так это о том, что вы случайно сделаете что-то не так и никто вам об этом не скажет, а узнаете вы это только через несколько месяцев, когда ваша (неправильная) работа уже будет пущена в оборот. Вот почему крайне важно задавать вопросы, пока вы находитесь в позиции новичка: так вы позже будете чувствовать уверенность в своих результатах. Постоянные неудачи приводят к быстрому выгоранию, поэтому работайте над созданием собственных процессов, которые позволят добиться успеха в долгосрочной перспективе.

9.1.3. Понимание и установка ожиданий

Одна из самых важных вещей, которую вы можете сделать в первые недели работы, — это встретиться со своим руководителем, чтобы обсудить приоритеты. Это важно, поскольку так у вас появится представление о том, над чем следует работать. В одних компаниях приоритетная задача — это предоставление анализа определенной группе стейкхолдеров, чтобы помочь им в развитии какой-либо

части бизнеса. В других — создание высокопроизводительных моделей для вебсайта. В некоторых организациях могут либо ставиться обе этих цели, либо не ставиться ни одна из них.

Вам может казаться, что после внимательного чтения вакансии и прохождения интервью вы уже знаете, чего следует ожидать от работы. И хотя порой так и есть, многое может измениться, прежде чем вы приступите к обязанностям. У интервьюеров может быть устаревшая информация или в компании могло что-то поменяться до вашего прихода. Поговорите с руководителем как можно раньше, чтобы получить наиболее актуальные данные и обсудить их при необходимости.

В идеале у вашего руководителя есть ви́дение того, чем вы будете заниматься, но он готов учитывать ваши приоритеты и сильные стороны. Вместе вы должны определить показатели эффективности. Как правило, ваш успех связан с успехом команды и/или руководителя; если не все сотрудники стараются для достижения одной и той же цели, поддерживать друг друга будет сложно. Чтобы определить, как добиться успеха самому, стоит понять, какие задачи стоят перед командой и как оценивается ее производительность. Поможете ли вы получить больший доход, работая над экспериментальными программами по увеличению конверсии, или создадите модель МО, чтобы помочь специалистам по обслуживанию клиентов спрогнозировать проблемы покупателей с целью сокращения среднего количества времени запроса?

«Создайте модель машинного обучения с точностью 99 %» или «Используйте новейшую статистическую модель в своем анализе» — обычно это *не* цели производительности. Это инструменты решения проблемы. Ваши модели и анализ бесполезны, если они не связаны с задачами, которые необходимо решить другим людям. Считать, что целью является разработка наиболее производительных моделей, — распространенное заблуждение среди тех, кто только начал свой путь в Data Science. Логично, что так думают многие, потому что большое число научных исследований и образовательных курсов сосредоточены на разных методах создания точных моделей. Но для достижения успеха в Data Science одного этого недостаточно. Такие параметры, как полезность модели, уровень получения аналитических оценок и удобство эксплуатации, часто бывают важнее. (В главах 10 и 11 мы подробнее остановимся на этом вопросе.)

Приступая к новой работе, вы не знаете наперед, какие ожидания связаны с вашими непосредственными обязанностями. Некоторые компании ценят командную работу; от вас могут ожидать работы над несколькими проектами одновременно, но при этом надеяться, что вы сразу же бросите свои задачи ради помощи коллеге. Другие организации ждут работы на результат и игнорирования сообщений в Slack. Чтобы узнать, насколько вы соответствуете ожиданиям,

нужно регулярно проводить совещания со своим непосредственным руководителем. В большинстве компаний у вас будут еженедельные встречи тет-а-тет для обсуждения текущих задач или вопросов. С их помощью можно узнать, уходит ли ваше время на важные для начальства задачи. Зачем тратить силы на догадки, если можно спросить прямо? Обдумывание краткосрочных блоков придаст уверенности, что вы на правильном пути, когда дело дойдет до более масштабной оценки эффективности.

Настройтесь на успех

В крупных компаниях обычно проводится официальная проверка эффективности сотрудников, поэтому обязательно спросите, что включает в себя этот процесс и когда это происходит. Часто это делается каждые шесть месяцев с возможным последующим повышением зарплаты и должности. Многие компании проводят 360-градусную аттестацию, в ходе которой вы получаете прямую обратную связь не только от руководителя, но и от коллег. Если это ваш случай, выясните, кто выбирает коллег для оценки — вы или ваш руководитель. Это нужно для того, чтобы вы знали влиятельных игроков.

У устоявшихся DS-команд может быть матрица оценки компетенций, показывающая, в каких областях вас будут рассматривать и чего ждать в зависимости от вашего стажа. Например, это могут быть специальные технические знания. От джуниора ждут только фундаментальных знаний, а еще он должен показать, что развивается в своей сфере; миддл может разбираться в одной предметной области, а синьор может быть экспертом широкого профиля, например отлично знать A/B-тестирование или работу с большими данными. Если такой матрицы нет, посмотрите, сможете ли вы вместе с руководителем выбрать для вас несколько направлений.

Запланируйте с начальством проверку ваших компетенций через три месяца, если такой практики в компании нет. Так вы будете знать, что находитесь на одной волне с руководителем, сможете совершенствоваться и планировать оставшуюся часть первых шести месяцев, а затем и года.

Смысл оценки эффективности не в том, что вам нужно преуспевать во всех сферах с самого начала. На самом деле большинство компаний не станет проводит формальную оценку результатов тех, кто проработал в компании менее полугода, потому что бóльшая часть этого времени была потрачена на адаптацию сотрудника. Это нужно скорее для того, чтобы у вас сложилась общая картина, когда вы определитесь со своей ролью и приступите к обязанностям.

9.1.4. Знайте данные, с которыми работаете

Конечно, нужно еще разобраться с частью, касающейся непосредственно Data Science. Если компания, в которую вы попали, уже занимается этим направлением какое-то время, начните с изучения отчетов сотрудников. Из них вы не только узнаете о том, какие типы данных хранит организация (и не только об основных

аналитических выводах), но и поймете, каким образом нужно предоставлять результаты проделанной работы. Бóльшая часть задач дата-сайентиста заключается в передаче информации сотрудникам без технических знаний, при чтении отчетов вы осознаете, что большинство ваших коллег — абсолютные гуманитарии. Обратите внимание на то, как авторы упрощают или усложняют какие-либо понятия. Учтите это, когда придет время составлять собственные отчеты.

Затем следует выяснить, где находятся данные, и получить к ним доступ. Для этого нужно знать, какая таблица их содержит и в какой системе они записаны. Возможно, наиболее часто используемые данные находятся в базе данных SQL, а все, что старше двух лет, расположено в HDFS (распределенной файловой системе Hadoop), для доступа к которой необходимо использовать другой язык.

Оцените данные, с которыми вы планируете регулярно работать, подходите к ним объективно. В некоторых таблицах есть документация (пакет с данными либо отчет о них), в которой объясняются возможные сложности с качеством или прочие особенности. Сначала прочтите эти документы, чтобы не пришлось изобретать велосипед. Затем бегло оцените несколько строк и сводную статистику. Если вы будете знать, что некоторые подписки начинаются позже или что в столбце часто отсутствуют значения, то избежите части проблем. Если вы нашли незадокументированные сюрпризы, лучший способ выяснить, что пошло не так, — спросить эксперта, занимавшегося этой таблицей. В крупной компании это может быть дата-сайентист или кто-то, кто занимался сбором данных. Вы можете обнаружить, что сюрприз — это реальная проблема, которую нужно исправить, или же такие вещи в работе вполне ожидаемы. Например, подписки, которые активируются позже, могут оказаться подписками, которые были приостановлены и возобновлены в конкретную дату. Или вы можете заметить, что купоны на прошлогоднюю новогоднюю акцию были использованы в мае этого года, и это произошло потому, что их выдала служба поддержки.

Одни компании предпочитают, чтобы данные для тестирования были отделены от реальных данных, тогда как другие объединяют их не задумываясь. В последнем случае вам следует спросить, нужно ли исключать определенные заказы или действия тестовых аккаунтов или особых деловых партнеров. Точно так же в некоторые датасеты входят пользователи, которые ведут себя совершенно иначе. Например, American Airlines однажды предложила безлимитный проездной, включающий билет для сопровождающего лица. Один из пассажиров использовал тариф для сопровождения незнакомцев, домашних животных и своей скрипки, а также мог летать несколько раз в день. Возможно, вы не столкнетесь со столь экстремальной ситуацией, но новые компании нередко предлагают сделки, которые впоследствии выглядят глупо (например, 10 лет доступа за $100) и которые, возможно, необходимо будет учесть в анализе.

Элин Фарнелл (Elin Farnell): мысли о переходе из научной сферы в индустрию

Проработав математиком в академической сфере восемь лет, я задумалась о переходе в индустриальный сектор, когда поняла, что наиболее ценные для меня аспекты работы занимают центральное место в коммерческих должностях, связанных с Data Science. Я сделала два исследовательских проекта, которые стали результатом сотрудничества с инжиниринговой компанией в рамках грантов Министерства обороны и Министерства энергетики. Больше всего в этих проектах мне понравилось, что наша исследовательская группа работала над интересными вопросами математики и при этом знала, что наши разработки будут использоваться для решения реальных задач. Я также оценила возможность изучить новый математический материал и сотрудничать с многопрофильной командой. Оказавшись в новой для меня среде, я заметила определенный контраст между новым и старым опытом:

- *Компромисс между шириной и глубиной*. В академических кругах главным приоритетом, особенно для начинающих исследователей, часто является выбор исследовательской программы, сосредоточенной на глубокой узкой подобласти. В индустрии, напротив, цель обычно состоит в решении широкого круга задач, что предполагает изучение и применение большого набора инструментов из любой сферы. Оба варианта могут нравиться по-разному. Степень проявления этого компромисса между шириной и глубиной меняется в зависимости от учреждения и области исследований в академических кругах или от направления, выбранного командой, и отраслевого проекта. Там, где ваши личные предпочтения соответствуют ширине, спектр глубины поможет оценить различные карьерные возможности.
- *Автономия*. С точки зрения выбора исследовательских проектов академическая среда предлагает значительную автономию. В индустрии же предполагается, что вы будете решать те задачи, которые поставит работодатель (обычно вам дается достаточно свободы выбора в плане способов решения). Как я сказала вначале, тут главное преимущество — это понимание, что то, над чем вы работаете, положительно повлияет на реальный мир. Также следует отметить, что существуют механизмы повышения автономии в индустрии; многие должности позволяют дата-сайентистам предлагать новые сферы для работы в будущем, а внутреннее или внешнее финансирование откроет возможности для новых проектов.
- *Баланс между работой и личной жизнью*. На основании личного опыта и по рассказам других людей я считаю, что баланс между работой и личной жизнью лучше в индустрии, чем в академической среде. Как правило, в научном мире довольно сложно установить границы, а брать работу на дом каждый вечер и в выходные — в порядке вещей. В корпорациях переработки тоже случаются, но в большей степени все зависит от сроков и, как правило, работа выполняется поэтапно. Баланс между работой и личной жизнью в значительно мере зависит от культуры в конкретном учреждении или компании, а также от того, как лично вы участвуете в этой культуре и вносите в нее свой вклад. Я знаю людей по обе стороны баррикад и видела примеры, когда человек успешно находил баланс и когда ему это совершенно не удавалось.

В процессе изучения данных вы выясняете их общую форму. Порой в небольшой фирме вам для начала придется поработать с инженерами, чтобы собрать больше информации, прежде чем общие данные станут подходящими. А в крупной организации вам придется расшифровывать десятки таблиц, чтобы понять, существует ли то, что вам нужно, в принципе. Возможно, вы ищете таблицы со столбцом «Заказ» в 12 базах. В идеале где-то должны быть сохранены документально подтвержденные и поддерживаемые таблицы основных показателей бизнеса, таких как транзакции или подписки. Но это, скорее всего, не относится к другим, менее важным датасетам, и вам следует попытаться узнать больше, если вы будете работать с плохо задокументированными сферами.

Убедитесь, что вы знаете, как данные к вам попали. Если вы работаете с чем-то вроде данных веб-сайта, они, скорее всего, будут проходить через несколько систем, чтобы попасть в БД, которую можно использовать. Каждая из этих систем наверняка каким-то образом изменяет данные. Если сбор внезапно остановился, вы должны знать, где искать проблему (а не впадать в панику!). Но в некоторых случаях данные вводятся вручную, например данные пациентов больницы или результаты опросов. В таких случаях не беспокойтесь о конвейерах, сосредоточьтесь на понимании многих атрибутов данных и потенциальных местах, где человек мог ввести их неправильно. Где бы вы ни оказались, вам практически всегда придется иметь дело с «грязными данными».

По мере разбора попробуйте записать любые проблемные места в данных и составить карту того, где и что находится. Такие факты сложно запоминать в процессе работы, а у многих компаний нет хорошей системы для документирования или обнаружения данных. Подобно тому как комментарии к коду помогают вам и другим людям понять его назначение в будущем, документирование данных приносит огромную пользу. Можно хранить эти файлы и на своем ноутбуке, но будет лучше, если у всех сотрудников компании будет к ним доступ. Так вы поможете новым сотрудникам компании и даже нынешним дата-сайентистам, которые не знакомы с какой-нибудь конкретной областью.

9.2. Становимся продуктивными

В конце концов, вы должны облегчать работу своего руководителя и снижать его нагрузку, но вначале все будет ровно наоборот, и это понятно. Чтобы начать работать на полную, потребуется больше времени, чем вы можете рассчитывать. Раздражение — это нормально, но помните, что в новой среде вы имеете дело с большой когнитивной нагрузкой. Вы пытаетесь уловить нормы (вероятно, негласные) того, сколько длится обед, во сколько начинается и заканчивается рабочий день, какие формы общения следует использовать, стоит ли закрывать

ноутбук, покидая рабочее место, и многое другое. Кроме того, вам еще нужно разобраться с целой системой данных.

Распространенное заблуждение — это идея, что вам нужно поскорее проявить себя: «Я должен делать все быстрее, иначе все будут удивляться, зачем меня вообще наняли». Это тот самый случай синдрома самозванца (глава 8). Если вы попали в адекватную компанию, вам дадут некоторое время на разгон. Вместо того чтобы работать быстро, позиционируйте себя как сотрудника, который будет приносить пользу в долгосрочной перспективе (в течение месяцев, а не недель). Вначале вы будете задавать больше вопросов («Могу ли я получить к этому доступ? Почему этот запрос выполняется так медленно?»), чем давать реальных ответов (в форме отчетов, анализов и моделей).

При этом вы все равно можете стать полезным раньше. Сосредоточьтесь на простых и полностью описательных вопросах, таких как: «Как распределяется количество наших клиентов?» или «Какой процент пользователей активен каждую неделю?» В процессе вы ознакомитесь с данными компании, а также столкнетесь с рядом сложностей, которые вас поджидают. Во время совещания с руководителем покажите часть своей незавершенной работы, чтобы понять, находитесь ли вы на правильном пути. Очень неприятно потратить много времени и обнаружить, что вы отвечаете не на тот вопрос, применяете метод, который ваш начальник ненавидит, или используете неправильный источник данных.

Сосредоточившись на более простых вопросах, вы не поставите себя в неловкое положение: например, не сделаете неверный вывод, отвечая на сложный вопрос, не изучив сначала подробно все данные. Порой возникают непростые ситуации, потому что если ваши заказчики не знакомы с Data Science, их первый вопрос может звучать примерно так: «Можете ли вы спрогнозировать, какие торговые сделки будут закрыты?» или «Как мы можем удержать максимальное количество пользователей?» В главе 12 мы поговорим о том, что одна из ваших задач как дата-сайентиста — вникать в бизнес-вопросы, чтобы преобразовать их в вопросы, связанные с данными. Если человек не знает или имеет неправильное представление об основных фактах (например, о том, какой процент пользователей совершает вторую покупку или сколько людей переходят по рекламному баннеру), то он не задаст правильные вопросы.

Две стратегии помогут вам стать продуктивным быстрее: задавайте вопросы и выстраивайте взаимоотношения. Задавая вопросы, вы быстрее поймете детали своей работы. Выстраивание взаимоотношений позволяет понять вашу роль в организации.

9.2.1. Задавайте вопросы

Наибольший тормоз вашей карьеры — это боязнь задавать вопросы или говорить: «Я не знаю». Как мы уже неоднократно упоминали, Data Science — это настолько

обширная область, что никто не знает всего или даже 20 % всего! И вы тоже никак не сможете узнать все тонкости данных в свой компании. Ваш руководитель предпочел бы, чтобы вы задали вопрос и тем самым отняли несколько минут чьего-то времени, чем застряли бы с решением на несколько дней. Вопросы могут быть какими угодно, от технических (например «Какой статистический тест мы используем для того, чтобы с помощью A/B-тестирования определить изменения дохода?») до связанных с бизнесом (например «Какая команда отвечает за этот продукт?»).

При этом не все вопросы равны. Вот несколько советов о том, как задать наиболее полезные вопросы:

- *Наблюдайте за культурой вопросов в компании*. Как их задают — лично, через корпоративный мессенджер Slack, на форуме или по электронной почте? Правильный выбор канала снижает вероятность того, что вы кого-то побеспокоите. Вы также можете уточнить этот момент у своего руководителя.
- *Проявите инициативу*. Вы можете сказать: «Я изучил этот вопрос и выявил три вещи» или «Это похоже на X. Это оно?» Проведя собственное исследование, вы, возможно, сумеете ответить самостоятельно и сможете задавать вопросы, лучше понимая концепцию.
- *Не задавайте вопросов, если можете быстро найти ответы самостоятельно*. Если вопрос не возник непосредственно в ходе разговора и если ответ выдается в первом же результате поиска в Гугле, лучше не задавайте его (например: «В чем разница между вектором и списком в R?»).
- *Найдите экспертов, которые могут вам помочь, уважайте их время*. Некоторые из ваших вопросов будут общими, но могут возникнуть и вопросы глубоко технического характера. Важно выяснить, кто является экспертом по различным методам статистики или программирования, поскольку именно от этих людей вам нужно получить ответы. Постарайтесь не становиться для них обузой: если вы понимаете, что вопросов к конкретному человеку накопилось много, попробуйте назначить с ним встречу. Людям гораздо проще потратить время на несколько организованных встреч, ограниченных по времени, чем отвечать каждые несколько минут. Поинтересуйтесь стилем работы своих коллег. Некоторые сотрудники должны помогать другим, но если от какого-нибудь человека также требуют результатов его непосредственной работы, посмотрите, есть ли у него календарь, где указаны часы, когда он недоступен. Чужое время следует уважать.
- *Не задавайте вопросы со скрытой критикой*: «Почему вы пишете код для этого запроса так, а не тем способом, которому я научился на последнем курсе?» Постарайтесь искренне понять, почему все делается именно так. Если компания существует на рынке уже какое-то время, у нее есть большой объем техниче-

ских работ по разным проектам. Например, если у крупной организации есть физические серверы, для переноса данных с них в облако потребуется уйма времени и усилий десятков инженеров. Когда люди спрашивают: «Почему бы нам просто не сделать X? Это так просто и сэкономит нам много времени», они часто предполагают, что другие не понимают проблемы или не считают ее срочной. Но причина, по которой компания не делает X, может быть связана с вещами, о которых вы не знаете, например из-за юридических ограничений.

- *Объединитесь с кем-нибудь еще*. Отличный способ учиться — объединяться с людьми. Вместо того чтобы просто спрашивать, можно проследить, как другие находят нужные ответы. Что касается технических вопросов, то объединение с кем-либо — это также способ увидеть среду программирования и изучить новые методы. Даже если ваш вопрос касается способа получения данных, можно узнать, в какой таблице они находятся и как человек нашел их там, а еще, возможно, получится освоить некоторые хитрости написания кода. Конечная цель — разобраться с как можно большим количеством вопросов, зная, где искать ответ.
- *Составьте список*. Если у вас есть вопросы, на которые не требуется ответить немедленно, запишите важные детали, которые вы хотели бы знать, например как часто обновляются данные, какой предельный размер для запросов и как далеко уходят определенные данные на локальном сервере. Затем обсудите все это вместе со своим наставником или руководителем. С таким подходом вам не придется постоянно отрывать кого-то от работы.

9.2.2. Выстраивайте взаимоотношения

Круг поддержки — это важная часть комфорта в новой рабочей среде. Одним его создание дается проще, чем другим; постарайтесь общаться с коллегами не только на технические темы. В большинстве случаев это означает организацию встреч с людьми, с которыми вы никогда раньше не общались, чтобы познакомиться с ними и их работой поближе. Это не время, потраченное напрасно: благодаря такому общению вы и ваши коллеги будете чувствовать себя комфортнее, имея возможность положиться друг на друга и знать больше, чем просто имена и должности.

Подойти к незнакомому человеку может быть непросто, но вы можете использовать вопросы как способ заговорить с кем-то. Людям нравится быть полезными и чувствовать себя компетентными, поэтому не бойтесь вежливо их спрашивать в знак дружеского внимания. Если вы знаете хотя бы несколько человек, то даже самые большие офисы не испугают вас. Нормальной практикой также считается написать людям, с которыми вы будете тесно взаимодействовать, назначить

с ними 30-минутную встречу, чтобы познакомиться. Если вы работаете в большом офисе, попросите вашего руководителя составить список сотрудников, которых вам стоит знать.

Наставничество и спонсорство

«Найдите наставника» — один из самых распространенных советов, но он может быть безнадежно непрактичным. *Наставник* или карьерный советник действительно может помочь в решении сложных задач и принятии правильных решений. Но в отличие от обучения программированию или улучшения коммуникативных навыков, наставничество не предполагает практических занятий для посещения или книг для изучения. Так как же его найти?

К счастью, наставничество не обязательно предполагает длительные взаимоотношения. Анджела Басса (Angela Bassa), с которой мы поговорим в главе 16, составила список людей, которые готовы отвечать на вопросы и наставлять новичков в Data Science на datahelpers.org. Наставник — это не обязательно тот, кому вы можете позвонить по поводу любой вашей карьерной дилеммы, зато к нему можно обратиться за помощью с конкретной проблемой, например если вы хотите потренироваться в поведенческом интервью или составить свой первый пакет на R.

Есть еще один тип людей, которые могут даже больше влиять на вашу карьеру: спонсоры. *Спонсор* — это тот, кто дает людям возможности, финансируя их инициативы, отстаивая их продвижение по службе, знакомя их с важными людьми или обеспечивая их назначение на те сложные проекты, которые могут помочь им расти профессионально. Даже больше, чем наставнику, вы должны доказать спонсору, что хорошо справитесь с предоставленной возможностью. Если кто-то порекомендовал вас в качестве докладчика на конференции, а вы ни разу не ответили организатору или выступаете с явно неподготовленным материалом, то это плохо отразится на вашей репутации и дальнейшем продвижении. Вам не обязательно предварительно иметь аналогичный опыт, но если вы покажете, что сделали нечто похожее (например, выступили на встрече), и если вы вежливы и исполнительны, то сможете убедить их, что сделаете свою работу хорошо.

Если вы хотите, чтобы кто-то стал вашим постоянным наставником или спонсором, расскажите этому человеку, как его совет или предоставленная им возможность вам помогли. Многие люди выступают в этой роли, потому что хотят помочь, и им будет приятно узнать о плодах своих трудов. Если же вы обращаетесь только когда вам что-то нужно, они могут решить, что вы просто их используете, а это никому не нравится.

Многие статьи о спонсорстве и наставничестве посвящены вопросу поиска этих людей в компании, что особенно актуально, если вы работаете в крупной организации. Но дата-сайентисты меняют работу каждые несколько лет, а возможностей находить спонсоров и наставников за пределами вашей компании, которые поддерживали бы вас даже в случае увольнения и перехода на новое место, не так много, так как эта сфера достаточно узкая.

В команде любого размера полезно узнать, к кому и с каким вопросом следует обращаться. Один сотрудник может быть лучшим специалистом компании по SQL, а другой может отвечать за систему экспериментов. Знать, к кому кроме руководителя можно обратиться в случае технических трудностей, может оказаться очень полезно. Также следует как минимум представиться начальнику вашего руководителя. Это нужно вовсе не для того, чтобы вы могли пожаловаться на него, а потому, что так им будет проще говорить о вас при необходимости.

Точно так же следует познакомиться со всеми участниками проекта, с которыми вы будете работать. Если в DS-команде меньше десяти человек, постарайтесь встретиться с каждым по отдельности. Если вы будете работать с дата-инженерами или другими специалистами по данным, поговорите с ними. Это может быть и неформальный разговор, но важно, чтобы о вас знали не только по подписи в имейлах. Даже если вы в основном работаете удаленно, попробуйте использовать систему видеоконференцсвязи, чтобы вас могли увидеть и запомнить ваше лицо.

Слушайте все, что говорят как на официальных совещаниях, так и во время обеда. Познакомьтесь с людьми, которые работают в областях, смежных с Data Science (от проектировщиков до финансистов и экспертов по продажам и маркетинговой аналитике), и узнайте больше о специфике их работы. Не торопитесь заявлять что-то вроде: «Я мог бы сделать это лучше» и не берите на себя преждевременные обязательства из серии «Мы создадим для вас платформу машинного обучения». Просто сосредоточьтесь на сборе информации. И помните, что не все разговоры должны быть завязаны на работе. Всегда хорошо спросить о планах на выходные, любимых телешоу или хобби.

Последний мудрый совет: подружитесь с офис-менеджером. Офис-менеджеры отвечают за множество вещей, которые могут сделать ваш день лучше, — закуски, заказ обеда, лосьон для рук в туалете и так далее. А еще у офис-менеджеров одна из самых тяжелых и неблагодарных работ, поэтому цените их.

9.3. Если вы первый дата-сайентист

Все, о чем мы говорили до этого момента, применимо к первым нескольким месяцам любого дата-сайентиста, но если вы стали первым специалистом по данным в компании, то здесь есть ряд уникальных проблем, с которыми вы неизбежно столкнетесь. Учитывая новизну этой области и нехватку кадров в небольших организациях, нередко случается так, что нанятый сотрудник становится первым. Поэтому, будучи первопроходцем, вы должны быть особенно подготовлены на старте.

Когда вы только приступите к работе, у вас не будет абсолютно никаких прецедентов. Еще никто не решал вопрос о том, стоит ли использовать Python, R или

какой-либо другой язык. Никто не придумал, как управлять процессом. Следует ли применять гибкие методологии разработки ПО, такие как Agile, чтобы расставлять приоритеты, или же работать по наитию? Как организовать код? Следует ли купить профессиональную лицензию GitHub, или использовать сервер Microsoft TFS, или же хранить все файлы в папке «Мои документы» на ноутбуке без резервного копирования?

Поскольку прецедентов в компании еще не было, им становится любое ваше действие. Например, если вы предпочитаете малоизвестный язык F#, то тем самым заставляете следующего дата-сайентиста его изучать. В ваших интересах принимать решения, которые принесут пользу независимо от того, какой будет команда в будущем, то есть это может означать выбор в пользу более распространенного языка, чем тот, который нравится лично вам. Но помните: во всем важен баланс и слишком большое внимание будущему может серьезно навредить настоящему. Если вы потратили три месяца на создание красивого конвейера для автоматического обмена отчетами с другими дата-сайентистами, но второй специалист за пять лет так и не появился, значит, ваше время потрачено впустую. Вам придется ежедневно принимать решения, которые будут иметь прямые или косвенные последствия.

Помимо того что вам нужно самостоятельно определить свои обязанности, вы еще должны продвигать Data Science среди остальной части организации. Поскольку раньше у компании не было таких специалистов, большинство людей не будет понимать, зачем вы здесь. Чем быстрее у сотрудников появится представление о вашей роли, тем выше вероятность того, что с вами захотят работать и оставить вас в компании. Сюда же относится и управление ожиданиями. Как мы уже говорили ранее, некоторые считают, что Data Science — это волшебная таблетка и что первый же специалист по данным сможет сиюминутно решить некоторые из наиболее серьезных проблем компании. Сформулируйте реалистичные ожидания относительно того, (а) что можно решить с помощью данных и (б) сколько времени потребуется на достижение этих целей. Так что в целом вам придется постоянно объяснять людям, что такое Data Science, а в частности — рассказывать о том, что вы можете сделать, чтобы помочь бизнесу. Если вы двадцатый дата-сайентист в команде, то, возможно, вы будете сидеть в углу и месяцами работать над моделями, но для первопроходцев этот вариант точно не прокатит.

Быть первым специалистом по данным всегда означает гораздо больше работы и рисков, но в то же время это приносит большие плоды. Принимая технические решения, можно выбирать то, которое вам больше по душе. Продвигая Data Science в организации, вы получаете больше влияния, вас все знают. А по мере роста команды вы становитесь ее руководителем, что может быть очень полезным для карьеры.

9.4. Если работа не соответствует обещанию

Устроившись на работу в Data Science и обнаружив, что она далека от ваших ожиданий, легко разочароваться. Теперь, попав в индустрию и проработав всего несколько месяцев, вам, возможно, придется начинать все сначала. Хуже того, вас может беспокоить мысль, что такой непродолжительный срок работы испортит ваше резюме. Значит ли это, что вам нужно терпеть и отработать хотя бы год? Работать в неблагоприятной обстановке, равно как и решиться на увольнение, — непросто. В следующих разделах мы рассмотрим две основные категории проблем, с которыми сталкиваются дата-сайентисты, — ужасная работа и токсичная рабочая среда. Идеального варианта решения этих проблем не существует, но мы опишем некоторые возможные стратегии смягчения последствий.

9.4.1. Ужасная работа

Во-первых, внимательно проанализируйте свои ожидания. Вас беспокоит что-то вроде «Все данные не очищены! Два дня я потратил только на их подготовку! Дата-инженеры не исправляют все сразу!» С такими проблемами вы будете сталкиваться повсеместно. Даже в крупнейших компаниях с сотнями инженеров дата-сайентисты сталкиваются с этим; данных так много, что их невозможно полностью проверить. Хотя основные таблицы должны быть чистыми и хорошо задокументированными, скорее всего, вам попадутся данные, которые придется улучшать или объединять с другими.

Один из способов проверить реалистичность своих ожиданий — узнать о работе у других специалистов по данным. Если вы закончили учебу с соответствующей степенью или прошли буткемп, спросите коллег или выпускников этого же направления, что они думают о среде, в которой вы работаете. Если у вас пока не такой широкий круг общения, сходите на митапы или вступите в виртуальное сообщество, если живете в маленьком городке или сельской местности. (Мы подробно рассмотрели эту тему в главе 5.) Если другие ваши коллеги из DS-команды прежде занимались Data Science в других компаниях, попросите их рассказать подробнее о том, как все устроено.

Другая проблема может заключаться в утомительных и скучных задачах. Например, изначально вас нанимали для прогнозирования, но на практике оказалось, что все ваши обязанности сводятся к нажатию раз в месяц кнопки повторного запуска на существующей модели. В этом случае посмотрите, сможете ли вы реализовать какие-то сторонние проекты в организации или автоматизировать какие-нибудь процессы. Если работа скучная, но не отнимает много времени, воспользуйтесь возможностью, чтобы заняться делами, связанными с Data Science. Продолжайте пополнять свое портфолио сторонними проектами, пишите по-

сты в блогах или проходите онлайн-курсы. Эта тактика поможет подготовиться к следующей должности.

В конце концов, учиться можно даже на плохой работе. Есть ли способ организовать рабочий процесс так, чтобы выполнять задачи, на которых можно было бы учиться? Какие области вы можете улучшить? Возможно, коллеги не помогут вам научиться писать более качественный код, но можете ли вы узнать, какие ошибки легко допустить при формировании DS-команды? Скорее всего, в вашей компании работают умные и доброжелательные люди, но тогда почему же все идет не так? Понимая, что конкретно вам не нравится, вы будете знать, на что следует обращать внимание при поиске работы в будущем, а также будете лучше подготовлены и сможете избежать ошибок, если когда-нибудь создадите собственную DS-команду.

9.4.2. Токсичная рабочая среда

В разделе 9.4.1 обсуждается плохая ситуация, но ее можно исправить. Но что, если работа действительно токсична? Что, если руководитель и стейкхолдеры имеют совершенно нереалистичные ожидания и угрожают вам увольнением, потому что вы не можете спрогнозировать пожизненную ценность клиента, так как у вас нет для этого данных? Или вас штрафуют, если ответы, которые вы даете, не соответствуют ожиданиям компании? Организации, которые плохо знакомы со сферой Data Science, могут ожидать, что вы решите основные проблемы компании, взмахнув волшебной палочкой. Вам могут сказать: «Постройте модель, чтобы определить, хорошо ли написан текст» — проблема, к решению которой никто в этой области даже близко не приступал. В этом случае следует скорректировать ожидания работодателя, иначе вы рискуете постоянно ощущать, будто не дорабатываете. В таких условиях сложно защищать себя, но обычно в любой компании работают умные люди. Если вам говорят: «Если бы вы были компетентнее, то смогли бы это сделать», это звоночек. Даже если бы с этой задачей мог справиться гораздо более опытный специалист, организация должна была учесть это при поиске сотрудника на эту должность.

Проблема может заключаться в отсутствии взаимодействия отдельных сотрудников и команд. Вместо взаимопомощи команды могут пытаться саботировать процессы друг друга. Они сосредоточиваются только на собственном прогрессе и могут даже рассматривать успех в компании как игру с нулевой суммой: если вы или ваша команда преуспеваете, значит, их команда проигрывает. Помимо нездоровой среды такая ситуация часто приводит к огромному количеству работы впустую, поскольку вы можете в итоге дублировать чужой проект лишь потому, что с вами не поделились данными.

Задача может вообще не иметь ничего общего с Data Science; среда может быть сексистской, расистской, гомофобной или враждебной в любом другом смысле.

Ежедневный поход на работу не должен вызывать чувство дискомфорта. Даже если к вам не проявляют открытую агрессию, перебивание на совещаниях, использование неправильных местоимений при обращении или вопросы в стиле: «Вы *вообще* откуда?» влияют на атмосферу.

К сожалению, подобные проблемы обычно решаются только при активном участии высшего руководства и всех остальных. Но токсичная рабочая среда часто указывает на отсутствие умения руководить или даже на то, что начальство активно способствует ее распространению. Если причина кроется в одном сотруднике, в идеале другие признают этот факт и человека уволят. Но если проблема носит массовый характер, ситуация может быть практически непоправима, а попытка изменить что-либо с позиции младшего сотрудника — быстрый рецепт выгорания. В таких ситуациях хорошо подумайте: возможно, вам стоит уйти.

9.4.3. Решение уволиться

Решение о том, стоит ли уйти с работы, — исключительно личное. Несмотря на то что никто не сможет дать вам простую инструкцию, которая сделает его безболезненным и легким, мы можем предложить несколько вопросов, которые помогут вам сделать правильный выбор:

- Достаточно ли у вас сбережений, есть ли у вашего партнера дополнительный источник дохода, который может вас поддержать, или члены семьи, у которых вы можете попросить в долг, если уйдете в никуда?
- Влияет ли эта работа на ваше здоровье или личную жизнь?
- Если проблема заключается в работе, обсуждали ли вы ее со своим руководителем и пытались ли найти решение?
- Можно ли перейти в другую команду или на другую должность, если не сейчас, то в течение нескольких месяцев?

Если, ответив на эти вопросы, вы поймете, что нужно уходить, то один из вариантов — немедленно начать искать другое место. Но, возможно, вас беспокоит, как будет выглядеть ваше резюме после столь непродолжительного срока работы и как это объяснить интервьюерам. Если вы проработали всего несколько недель, а до этого уволились из предыдущей компании, подумайте о том, чтобы связаться с прежним руководителем. Скорее всего, на вашу должность еще не взяли нового сотрудника, и если вы расстались на хорошей ноте, то сможете вернуться.

Если вы ищете новую работу, вот несколько советов, как рассказать о своем непродолжительном опыте на собеседовании:

- *Подождите, пока интервьюер поднимет эту тему.* Не думайте, что об этом нужно говорить заранее; возможно, на это даже не обратят внимания, потому что компания явно заинтересована. (Вы дошли до стадии интервью!)

- *Расскажите о положительном опыте и извлеченных уроках*. Это могут быть ваши проекты, знакомство с отраслью или опыт, который вы переняли от руководителя.
- *Когда вас спросят, почему вы так быстро уволились, отвечайте коротко и нейтрально*. Вы находитесь в сложной ситуации, потому что должны честно рассказать о причинах ухода и о том, что это произошло не по вашей вине, но если вы будете чересчур откровенничать, интервьюер может несправедливо посчитать вас человеком со сложным характером. Поэтому лучше всего дать размытый ответ: «Требования к моей работе были не такими, как я ожидал, и я не смог применить свои навыки и знания на благо компании» и остановиться на этом. Если вы поняли, какая рабочая среда вам нужна, скажите об этом. Возможно, вы были первым дата-сайентистом в компании и осознали, что хотите стать частью более крупной команды. Если вы решили уволиться, ознакомьтесь с главой 15, где говорится о том, как сделать это достойно, а также даются советы о поиске нового места при работе на полную ставку и способах уйти на хорошей ноте.

Но, возможно, вы не можете уволиться, потому что ваша виза привязана к месту работы или вы работаете в единственной компании, которая занимается Data Science в вашем маленьком городке. Если это про вас, то вот несколько советов:

- *Помните, что вы — не ваша работа*. Вы не обязаны нести ответственность за неправильные решения, которые принимаются в компании. Если вы не руководитель, то, скорее всего, мало влияете на деятельность компании.
- *Берегите здоровье*. Не жертвуйте сном, спортом и временем с друзьями и семьей.
- *Поговорите с кем-нибудь*. Это может быть ваш партнер, друг или терапевт. Они могут дать совет, но иногда достаточно просто выговориться.
- *Подумайте о том, чтобы рассказать о харассменте в вашу сторону*. Если это делает конкретный человек, рассмотрите вариант о том, чтобы сообщить о нем в отдел кадров. Убедитесь, что ваше обращение задокументировано. Не обращайтесь в отдел кадров лично; отправляйте имейлы и получайте обратную связь, чтобы подтвердить обращение. Рано или поздно вы захотите сослаться на обидные слова, сказанные в вашу сторону, и наличие письменного подтверждения пригодится. Если компания ничего не делает, вы можете подать жалобу в Комиссию по соблюдению равноправия при трудоустройстве, если вы находитесь в США. К сожалению, подобного рода жалобы несут в себе определенный риск: несмотря на то что это незаконно, компании могут мстить сотрудникам, препятствовать их карьерному росту или даже уволить. Если вы все же не хотите писать жалобу, постарайтесь сохранить доказательства о любом харассменте, с которым вы столкнулись, на случай, если позже вы все же решите сообщить об этом.

- *Подумайте насчет увольнения нестандартно*. Может быть, вы думаете, что не можете уволиться, потому что единственные доступные вам варианты — переход в менее престижную компанию, более низкая должность или сокращение дохода. Наш вам совет: не стоит недооценивать негативные последствия пребывания в токсичной среде; если, уволившись, можно принести временную жертву, в долгосрочной перспективе оно того стоит.

Мы надеемся, что вы никогда не попадете в такую ситуацию, но лучше иметь где-нибудь в запасе информацию, которой вы сможете воспользоваться в случае чрезвычайной ситуации. Имейте в виду, что смена места работы в Data Science — обычное дело (мы обсудим это в главе 15), поэтому нет никаких причин держаться за компанию, где вам некомфортно.

9.5. Интервью с Джарвисом Миллером, дата-сайентистом в Spotify

Джарвис Миллер (Jarvis Miller) — дата-сайентист в Spotify, где он занимается улучшением аудиовпечатлений каждого пользователя. Когда мы брали это интервью, он работал специалистом по данным в BuzzFeed. Джарвис получил высшее образование в 2018 году со степенью магистра статистики.

Что вас удивило на первом месте работы в Data Science?

Меня удивили две вещи: то, насколько я могу стать лучше как писатель, и что мне нужно объяснять, чем направление Data Science полезно для бизнеса, не используя при этом жаргон. Я всегда считал, что поскольку стекхолдеры работали со специалистами по данным, то они научились понимать язык друг друга и мне не придется делать это иначе. Я осознал, что это абсолютно не так и нельзя просто сказать: «Я провел логистическую регрессию на этих данных для классификации…» Что касается моего умения лучше писать, я научился выражать суть, когда создаю отчет; я стал лучше излагать свои мысли и объяснять так, чтобы продакт-менеджеры, дизайнеры и стейкхолдеры, вообще не разбирающиеся в технологиях, понимали меня.

Я пришел из академической среды, и тогда мне казалось, что самое главное — это получение результата в конце дня; и не важно, начали ли вы работать непосредственно перед дедлайном или спланировали все заранее. В индустрии перед вами стоит большая общая цель, но вы разбираетесь, как разбить ее на версии. Вы запускаете первую версию, загружаете ее, узнаете, хорошо ли она работает или нет, и при необходимости улучшаете ее в следующем квартале. Я привык работать до тех пор, пока задача не решена. Но здесь мне пришлось научиться расставлять приоритеты между разными частями проекта, а затем завершать его. Я научился документально фиксировать выполненную мной работу, а также отмечать, что

нужно сделать в следующей версии; еще я стал делать результат своей работы общедоступным, будь то размещение отчета в общей папке или создание приложения, чтобы люди могли использовать мою работу и видеть, для чего она нужна.

С какими проблемами вы столкнулись?

Мне было сложно говорить открыто. Когда я пришел в компанию, мне поручили отдельный проект, а человек, у которого я был в подчинении, работал в Нью-Йорке, при том что я находился в Лос-Анджелесе. Если у меня возникал вопрос, я не знал, нужно ли сразу его задать или же отложить до совещания. Я понимал, что не хочу, чтобы что-то мешало моей работе, но при этом был не уверен, что можно считать проблемой, а что нет. Я думаю, что это распространенная ситуация, особенно для тех, кто принадлежит к маргинализированным группам или перешел из другой области. Такие люди могут чувствовать, что раз они новички и не эксперты, то не могут выразить недовольство или высказать свое мнение. Если бы я мог вернуться в прошлое, то, скорее, обсудил бы вопрос о собственном чувстве изоляции и непонимании того, как в этой компании устроена коммуникация.

Расскажите об одном из ваших первых проектов.

Одним из них было обновление платформы A/B-тестирования — очень обширная задача. Я начал с того, что составил список людей, с которыми можно было поговорить об их обязанностях в BuzzFeed и о том, на что похожа их работа и как A/B-тестирование в нее вписывается. Затем мы обсудили конкретный инструмент: что этим людям в нем не понравилось и почему, а также как он влиял на рабочий процесс. К сожалению, это привело к тому, что мне пришлось брать на себя слишком много всего. У большинства людей было стразу несколько предложений, и я учитывал каждое в равной степени, что в итоге вылилось в 50 важных задач, поставленных передо мной. Мой руководитель попросил меня разбить эти предложения на две группы — «обязательные» и «не помешает»; кроме того, нужно было аргументировать, почему я решил именно так. Он предложил задать общую цель проекта и расставить идеи по приоритету, исходя из их вклада в достижение цели и того, сколько времени потребуется на их реализацию.

Что бы вы посоветовали дата-сайентистам в их первые несколько месяцев работы в компании?

Помните, что вас наняли не просто так: вашу точку зрения уважают и считают, что помогут вам узнать что-то новое, а также уверены, что у вас тоже есть чему поучиться. Делитесь своим мнением. Если вы не любите выступать публично, отправьте сообщение одному человеку, обсудите вопрос с ним и обменяйтесь идеями, прежде чем выразить их перед большой группой.

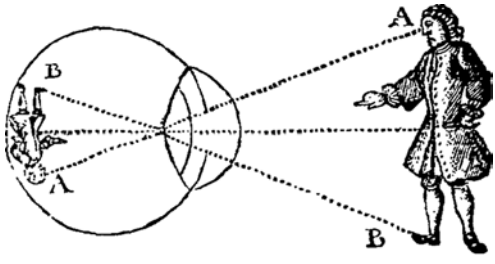
Это относится не только к техническим вопросам. Лично я предпочитаю не начинать общение с руководителем с вопросов о проделанной мной работе. Мне

нужно пару минут непринужденной беседы, чтобы я мог расслабиться, настроиться и освободиться от лишних мыслей. Я понимаю, что это делает меня продуктивнее, а работодателю нужен продуктивный сотрудник, так что ему стоит сказать об этой моей особенности. Ваше мнение ценится, и им стоит поделиться, особенно если речь идет о предпочтительном к вам отношении или о том, в каких условиях вы наиболее продуктивны. Другие знают вас гораздо хуже, чем вы сами себя, а ваша продуктивность будет всем на пользу.

Итоги

- Не пытайтесь стать продуктивным сразу. Сосредоточьтесь на выстраивании взаимоотношений, тактике работы и понимании данных: все это сделает вас продуктивным в долгосрочной перспективе.
- Если вы оказались в неблагоприятной рабочей среде, постарайтесь контролировать ситуацию, чтобы смягчить последствия для вашего здоровья и карьеры.

10



Создание эффективного анализа

В этой главе

- Планирование анализа.
- Продумывание кода, данных и структуры проекта.
- Передача анализа клиенту.

Эта глава написана для дата-сайентистов, которые специализируются в области принятии решений и аналитики, — для тех людей, которые предлагают компаниям идеи на основании оценки данных. Несмотря на то что инженеры МО также должны производить анализ перед тем, как создавать и развертывать модели, часть контента, связанного с менеджментом стейкхолдеров и красивой визуализацией, не всегда относится к их задачам. Если вы инженер по машинному обучению и читаете эту книгу, не спешите пропускать эту главу — она для вас по-прежнему актуальна; глава 11 вам тоже понравится: в ней рассказывается о развертывании моделей в производство.

В основе многих должностей в Data Science лежит *анализ*, включающий составление кратких документов, которые с помощью данных объясняют бизнес-ситуации или предлагают варианты решения коммерческих задач. Современные компании построены на отчетности и анализе. Тем, кто принимает решения, неудобно работать без данных, подтверждающих их выбор, а дата-сайентисты — это одни из лучших специалистов, которые понимают эти данные и находят в них смысл.

Анализ также важен для создания инструментов МО, потому что перед построением модели необходимо понять контекст датасета. Провести анализ, с помощью которого можно подытожить огромный объем данных компании, и тем самым пролить свет на рассматриваемый вопрос, чрезвычайно сложно и практически приравнивается к искусству. Разве можно ожидать, что человек возьмет таблицы с миллионами записей о прошлых периодах, каждая со своими сложностями и нюансами, и превратит все это в однозначное: «Да, данные говорят, что идея хороша»? Выяснение значимых с математической точки зрения деталей, вычленение важных для бизнеса моментов и поиск способов все это объединить — непростая задача, которая не решается просто так естественным образом.

В этой главе мы рассмотрим основы проведения анализа, чтобы вы понимали, какие результаты значимы для компании. Используя навыки, описанные в главе, вы сможете быстрее развиваться и двигаться вверх по карьерной лестнице.

Отчетность против проведения анализа

Отчет и анализ похожи, но это не одно и то же. Отчеты обычно создаются регулярно без значительных структурных изменений между версиями. Например, ежемесячный финансовый отчет может выглядеть как большая электронная таблица Excel, в которой каждый месяц обновляются цифры. Цель отчета — держать людей в курсе изменения показателей. Анализ делается один раз, чтобы ответить на более глубокий вопрос. Анализ привлечения клиентов, который показывает, как новые покупатели приобретают продукты, может быть проведен в R, а его результаты представлены в презентации PowerPoint. Отчеты, как правило, наполнены числами и показателями, тогда как анализ сосредоточен на предоставлении единственного основного результата. Большинство признаков хорошего анализа применимы и к отчетам, поэтому в этой главе мы используем термин «анализ» как универсальный, если не указано иное.

Что такое аналитический материал на самом деле? Как правило, он представляет собой слайды PowerPoint, файл PDF или Word или электронную таблицу Excel, которая может быть предоставлена сторонним специалистам и содержит информацию из данных и визуализаций, которые их отображают. На рис. 10.1 показан пример слайда, который можно найти в аналитическом материале. На анализ обычно уходит от одной до четырех недель, когда дата-сайентист должен собрать данные, запустить на их основе код для статистических методов и получить окончательный результат. После завершения этого процесса код не изменяется до проведения следующего анализа или, возможно, остается таким навсегда. Вот несколько примеров анализа:

- Анализ результатов опроса клиентов, который проводится с целью определить, какие продукты нравятся пользователям больше всего.

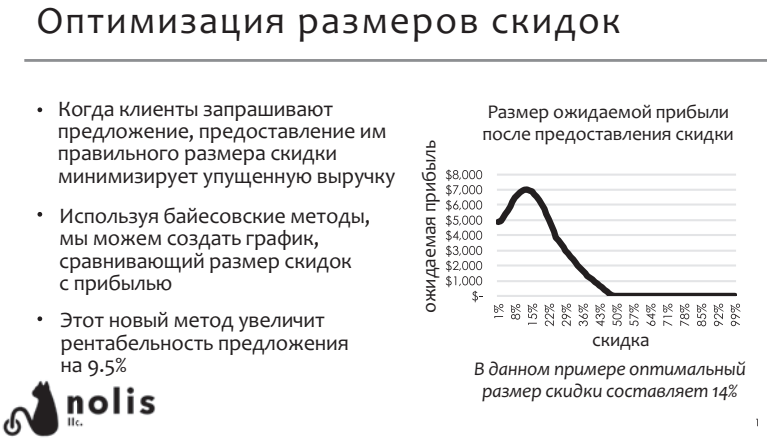


Рис. 10.1. Пример слайда PowerPoint с аналитическим материалом

- Анализ информации о местоположении, откуда делаются заказы. Проводится для выбора местоположения нового завода.
- Анализ архивных данных авиационной отрасли, который нужен для прогноза увеличения числа рейсов в определенные города.

Эти примеры имеют разный уровень технической сложности; некоторые требуют только обобщения и визуализации данных, тогда как для других нужны методы оптимизации или модели МО, но все они отвечают на один вопрос.

Так что же делает хороший анализ? Ниже приводятся пять признаков:

- *Отвечает на вопрос*. Анализ начинается с вопроса, поэтому, чтобы считаться содержательным, он должен давать ответ. На вопрос: «На каком из этих двух веб-сайтов клиенты покупают больше товаров?» анализ должен показать, на каком веб-сайте больше продаж. Ответ может быть даже такой: «У нас недостаточно информации, чтобы сказать точно», но это должен быть прямой ответ на вопрос.
- *Сделан быстро*. Ответы на вопросы, связанные с бизнесом, повлияют на решения, у которых есть сроки. Если на создание анализа уходит слишком много времени, решение будет принято без него. Обычно предполагается, что результат будет в течение месяца.
- *Им можно поделиться*. Анализ должен быть передан не только его заказчику, но и тем людям, с которыми этот человек хочет им поделиться. Например, если анализ содержит график, нельзя оставить его внутри скрипта R или Python; он должен быть в читаемом общедоступном формате, например PowerPoint.

- *Понятный*. Поскольку вы заранее не знаете, кто будет смотреть анализ, он должен быть понятен сам по себе. У графиков и таблиц должны быть четкие описания, помеченные оси и записанные пояснения; по возможности в нем не должно быть ссылок на другие работы.
- *Можно к нему вернуться*. Большинство вопросов возникнет снова. В некоторых случаях это означает повторное выполнение точно такой же работы, например вам придется еще раз провести кластеризацию. Иногда аналогичный подход нужно использовать где-то еще, скажем если нужно изменить входные данные и вместо европейских клиентов указать азиатских.

Эти черты приводят к основной мысли: «Хороший анализ — это то, что помогает специалистам, не занимающимся данными, делать свою работу».

Остальная часть этой главы выстроена так, чтобы охватить этапы анализа в хронологическом порядке, начиная с первоначального запроса и заканчивая передачей отчетов. Большинство видов анализа будут (или должны) соответствовать этой хронологии, но не все. По мере того как вы набираетесь опыта в этой работе, у вас может возникнуть искушение пропустить некоторые шаги, но именно такое поведение является причиной ошибок, которые допускают специалисты по данным.

Анализ для разных типов дата-сайентистов

Ситуации, в которых вы будете заниматься анализом, будут в значительно мере зависеть от ваших должности и обязанностей:

- *Специалист, принимающий решения*. В этом случае анализ является основной обязанностью. Специалисты по принятию решений постоянно углубляются в данные, чтобы ответить на вопросы, которые необходимо донести до бизнеса. Анализ — ключевой инструмент.
- *Инженер по машинному обучению*. Хотя инженер МО занимается созданием и развертыванием моделей, анализ по-прежнему является полезным инструментом для оценки качества их работы. Анализ нужен, чтобы показать ценность построения новых моделей или то, как они со временем меняются.
- *Аналитик*. Аналитики — это те дата-сайентисты, которые уделяют большое внимание показателям и KPI бизнеса и обычно делают множество отчетов. Они создают поток повторяющихся данных для компании, часто в Excel, SQL, R или Python. Хотя эти эксперты будут проводить анализ, им нужно думать об удобстве сопровождения работы больше, чем другим специалистам, потому что им часто приходится повторять ее.

10.1. Запрос

Анализ начинается с того, что вам отправляют запрос с требованием ответить на бизнес-вопрос. Человек из другого отдела или ваш руководитель спросит вас о чем-

то вроде: «Можете ли вы узнать, почему виджеты в декабре плохо продавались в Европе?» или «Отличается ли поведение среди наших клиентов малого бизнеса и представителей крупных организаций?» В зависимости от уровня технической подготовки этого человека можно получить либо неточный («Почему продажи падают?») либо точный запрос («Какие атрибуты коррелируют с более низкой средней стоимостью заказа?»).

Анализ строится вокруг бизнес-вопроса, но Data Science не строится на его основе. DS-вопросы могут звучать так: «Как нам кластеризировать эти точки данных?» или «Как нам спрогнозировать продажи?» Дата-сайентист должен преобразовать бизнес-вопрос в DS-вопрос, ответить на него и преобразовать в язык бизнеса. Это непросто. Чтобы понимать взаимосвязь этих двух видов вопросов, требуется опыт взаимодействия с типом рассматриваемой проблемы. Кроме того, необходимо понимать потенциальную полезность результатов различных статистических методов. Этот метод трансформации бизнес-вопросов в DS-вопросы и обратно был разработан Рене Теате, интервью с которой вы найдете в главе 14.

Этот процесс показан на рис. 10.2. Бизнес-вопрос поступает от стейкхолдеров, которые хотят знать, как таргетировать маркетинг под разных клиентов. Дата-сайентист должен перевести этот вопрос на язык математики — в этом примере это кластеризация данных о клиентах. Когда процесс завершается, у специалиста есть ответ (например, набор из трех групп кластеризованных точек данных). Наконец, дата-сайентист должен перевести ответ обратно на тот язык, который будет понятен бизнесу, например обозначить группы «Новые клиенты» или «Люди с высокими расходами».

Рис. 10.2. Процесс поиска ответа на вопрос бизнеса с помощью Data Science, предложенный Рене Теате

Прежде чем приступить к изучению данных и написанию кода, необходимо проделать подготовительную работу, чтобы лучше понять бизнес-вопрос. Чтобы предоставить полезный результат, нужно понимать контекст анализа. Кто просит провести его и какое отношение имеет этот человек к команде? Какую цель

он преследует? Это конкретный вопрос, на который нужен такой же ответ, или же расплывчатое общее представление о проблеме и надежда, что данные могут внести ясность? А есть ли у вас вообще данные для решения этой задачи? Если нет, то что нужно, чтобы их получить? Задавая вопросы, вы не только определяетесь со способом решения, но и понимаете, для чего оно будет использоваться. Многие специалисты по данным потратили недели на анализ и лишь позже обнаружили, что стейкхолдеру было «просто интересно», а в работе не было никакой необходимости.

На эти вопросы обычно отвечают на 30–60-минутном совещании с заказчиком и остальными участниками процесса. Вы будете заниматься непосредственно анализом и, возможно, не будете отвечать за организацию такого совещания. Однако если в вашем календаре эта встреча не появилась, стоит ее запланировать. Если вы не знакомы с человеком, который запрашивал анализ, это отличный шанс узнать о его работе.

Гипотетический пример набора фундаментальных знаний будет примерно таким:

- *Кто запрашивает анализ*? Юлия из команды разработчиков виджетов.
- *Для чего он нужен*? В этом месяце продажи виджетов упали на 10 %, и бизнес-команда не знает причину.
- *Что нужно выяснить*? Команда хочет использовать данные, чтобы понять, было ли падение характерно только для одной части страны.
- *Какое решение будет принято*? Решение состоит в том, следует ли прекратить выпуск этого виджета или нет.
- *Есть ли у нас необходимые данные*? Да, для анализа нужны заказы клиентов по почтовому индексу доставки, который доступен в БД заказов.

Очень важно знать, есть ли у вас данные, чтобы достоверно ответить на вопрос. Наверняка вам бы не хотелось потратить несколько недель, только чтобы в итоге вернуться к стейкхолдеру с пустыми руками.

Пример ситуации, в которой у вас нет данных, выглядит как-то так: в розничной компании хотят узнать, сколько заказов сделал каждый покупатель, но поскольку оплата осуществляется наличными, то использовать существующие данные для персонализации нереально. В такой ситуации лучше всего честно сказать заказчику, что выполнить его просьбу невозможно. Вам могут предложить альтернативные способы использования данных, которые могут быть достаточно близкими к тому, на что вы рассчитывали, или, возможно, вам придется объяснить, почему альтернативы тоже не работают. Если это возможно, предложите план получения необходимых данных. В предыдущем примере программа лояльности позволила бы связать заказы с конкретным клиентом и таким образом решить проблему с данными, но для создания такой программы потребуется время.

Другие вопросы из разряда: «Кто этот человек и почему он делает запрос?» полезны для создания плана анализа.

10.2. План анализа

Для дата-сайентистов нет ничего увлекательнее, чем погрузиться в данные, чтобы ответить на вопросы. Загружаем данные! Группируем! Подводим итог! Подбираем модель и получаем результаты! К сожалению, из-за бесчисленных способов обобщения и моделирования данных может получиться так, что вы потратите недели на работу и лишь потом обнаружите, что ничего из созданного вами не отвечает на поставленный бизнес-вопрос. *Худшее* из ощущений — осознание, что ничего полезного сделать не получилось. С аналитиками такое часто случается, особенно с джунами без солидного опыта.

Один из способов решить эту проблему — установить рамки, чтобы держать все под контролем и выполнять только нужную работу. План анализа — это и есть рамки. Идея заключается в том, что перед просмотром данных вы записываете все, что планируете с ними делать. Затем, по мере продвижения анализа, вы проверяете, какая часть плана выполнена. Когда все его пункты выполнены, работа завершена! Так можно не только сверяться с планом, но и отслеживать прогресс и обеспечивать отчетность. Ее можно даже показать на совещании с руководителем, чтобы обсудить, как идут дела.

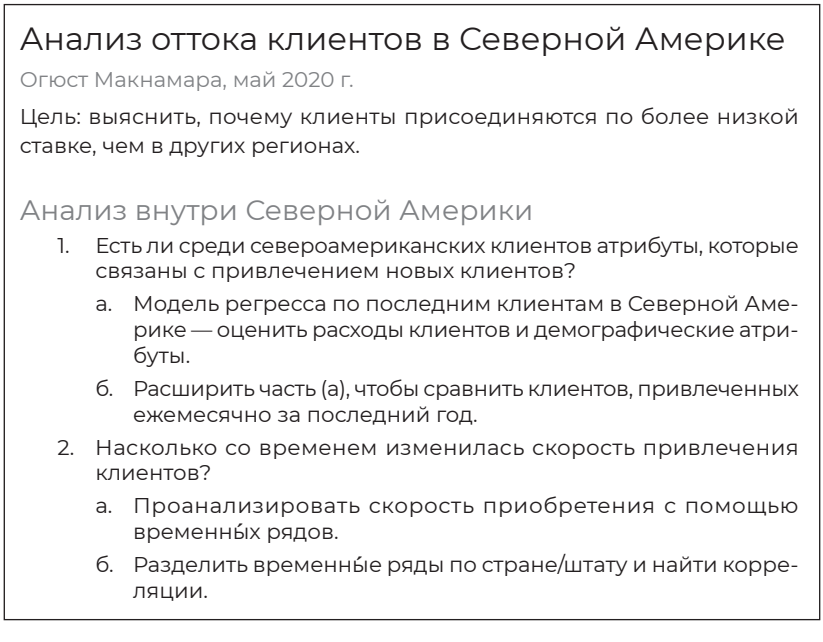
При составлении плана все его пункты должны быть выполнимыми. Можно написать код для задачи «сделать линейную регрессию продаж по регионам», но у вас не получится сделать то же самое для формулировки «выяснить, почему продажи упали»; это результат других вещей. Если задачи в плане выполнимы, оценить прогресс легко. Это также упростит анализ, потому что вам не придется ломать голову над дальнейшими действиями. Вместо этого вы просто смотрите на план анализа и выбираете следующую задачу.

Чтобы вам было легче составить несколько первых планов анализа, мы настоятельно рекомендуем использовать следующий шаблон:

- *Начало.* Укажите название анализа, его цель и ваши данные (в случае, если доступ к нему будет у кого-то еще).
- *Разделы.* Каждый раздел должен представлять общую тему анализа. Он должен быть понятным и независимым (анализ не полагается на результаты других разделов), чтобы другой человек мог работать над любым из них. В каждом разделе должен быть список задач.
- *Первый уровень раздела.* Здесь должен быть заданный вопрос. Он напомнит остальным, почему вы выполняете эту задачу. Если вам удастся найти правильные ответы, значит, вы поняли тему основного раздела.

- *Второй уровень раздела*. Здесь должны быть указаны задачи, которые можно отмечать по мере выполнения работы. Например, можно указать типы моделей для запуска; описания должны быть достаточно конкретными, чтобы в любое время можно было точно определить статус завершения работы.

На рис. 10.3 показан пример плана анализа. У нас это план оценки причин, по которым клиенты уезжают из Североамериканского региона. Вверху указаны название, цель и контактная информация дата-сайентиста на случай, если этот материал будет кому-то передаваться. Каждый раздел охватывает отдельный компонент анализа (например, анализ по Северной Америке или сравнение с другими регионами). Подразделы (пронумерованные) — это вопросы анализа, а самый нижний раздел (отмеченный буквами) — это конкретные задачи, которые необходимо решить.



Анализ оттока клиентов в Северной Америке

Огюст Макнамара, май 2020 г.

Цель: выяснить, почему клиенты присоединяются по более низкой ставке, чем в других регионах.

Анализ внутри Северной Америки

1. Есть ли среди североамериканских клиентов атрибуты, которые связаны с привлечением новых клиентов?
   а. Модель регресса по последним клиентам в Северной Америке — оценить расходы клиентов и демографические атрибуты.
   б. Расширить часть (а), чтобы сравнить клиентов, привлеченных ежемесячно за последний год.
2. Насколько со временем изменилась скорость привлечения клиентов?
   а. Проанализировать скорость приобретения с помощью временны́х рядов.
   б. Разделить временны́е ряды по стране/штату и найти корреляции.

Рис. 10.3. Пример плана анализа

При составлении плана поделитесь им с руководителем и стейкхолдером, сделавшим запрос. Они должны либо предложить варианты по его улучшению, либо одобрить его. Утвержденный план анализа обеспечивает согласованную базу для работы. Если по итогу вас спросят, почему вы так поступили, можете сослаться на утвержденный документ с исходными целями.

Вполне вероятно, что в процессе работы вы поймете, что упустили что-то важное, или у вас появится новая идея. Это совершенно нормально; просто

обновите план и сообщите стейкхолдеру, что вы вносите изменения. Поскольку ваше время ограничено, возможно, придется пожертвовать какой-нибудь менее важной задачей. План анализа полезен, потому что он создает диалог вокруг удаления ненужных пунктов, чтобы вам не приходилось пытаться выполнить невозможный объем работы.

10.3. Выполнение анализа

С подписанным планом на руках вы можете приступать к работе! Для начала производится импорт данных для дальнейших очистки и обработки. Затем вы несколько раз преобразуете данные путем их обобщения, агрегирования, изменения, визуализации и моделирования. Когда данные готовы, вы передаете их другим специалистам.

В следующих разделах мы кратко рассмотрим некоторые вопросы, которые следует учитывать при выполнении таких задач в рабочей среде. Целые книги, посвященные этой теме, также могут научить вас писать код для анализа на выбранном вами языке.

10.3.1. Импорт и очистка данных

Прежде чем вы сможете начать отвечать на вопросы согласно плану анализа, вам необходимо сохранить данные в удобном формате там, где с ними можно будет работать. Обычно это означает возможность загрузить их на R или Python, но иногда используется SQL или другие языки. Почти всегда эта задача отнимает у вас больше времени, чем вы рассчитываете. В процессе может возникнуть много сюрпризов. Вот некоторые из этих ужасных моментов:

- Проблемы с подключением к базам данных компании в конкретной интегрированной среде разработки (IDE).
- Проблемы с неправильными типами данных (например, числами в виде строк).
- Проблемы со странными форматами времени («год-день-месяц» вместо «год-месяц-день»).
- Данные, требующие форматирования (возможно, каждый идентификатор заказа начинается с элемента «ID-», который необходимо удалить).
- Отсутствующие записи.

Хуже того, ни одна из этих задач не выглядит продуктивной для людей, далеких от технических вопросов; вы не можете показать стейкхолдеру убедительный график работы драйвера БД, да и он не поймет, что манипуляции со строками помогают решить его бизнес-вопрос. Поэтому, какой бы утомительной ни была эта задача, вам нужно ее поскорее решить и приступить наконец к исследованию данных.

Занимаясь импортом и приведением данных в порядок, помните, что перед вами стоит двойная задача: побыстрее решить не столь важные моменты и уделить как можно больше времени задачам, которые помогут в дальнейшем. Если у вас есть столбец с датами, который представлен в виде строк, и вы сомневаетесь, что он вам когда-либо понадобится, не тратьте время на изменение формата. Но если вы считаете, что он пригодится, сделайте эту работу как можно скорее, потому что для анализа нужен чистый датасет. Трудно сказать заранее, что пригодится, а что нет, но если вы заметили, что времени на определенную задачу уходит слишком много, спросите себя, действительно ли она так нужна.

При импорте и очистке данных какая-нибудь проблема может выбить вас из колеи на несколько дней, например подключение к БД. Если вы оказались в такой ситуации, у вас есть три варианта: (1) попросить о помощи, (2) найти способ полностью избежать проблемы или (3) продолжать попытки решить ее самостоятельно. Вариант (1) подходит отлично, если вы можете к нему прибегнуть: человек, разбирающийся в вопросе лучше вас, может быстро найти решение, а вы сможете у него поучиться. Вариант (2) тоже хорош — можно, например, использовать файл .csv вместо подключения к базе. Варианта (3) — бесконечных попыток — стоит избегать любой ценой. Сотрудник, потративший несколько дней на одну проблему, выглядит бесполезным. Если какая-то задача ставит вас в тупик, обсудите с руководителем дальнейшие действия; не стоит продолжать попытки и надеяться, что проблема решится сама собой.

После загрузки и форматирования можете приступать к работе и искать *неадекватные данные*. Сюда относится все, что выходит за рамки фундаментальных представлений. Например, если при просмотре архивных данных о рейсах авиакомпаний обнаружилось, что несколько бортов приземлились до взлета, это странно, потому что обычно самолеты сначала взлетают! Выбиваться из ряда может что угодно, от магазина, продающего товары с отрицательной ценой, до производственных данных, показывающих, что на одном заводе произвели в тысячу раз больше товаров, чем на аналогичном. Такие странности наблюдаются постоянно, и их невозможно спрогнозировать до проверки.

Если вам встретились такие данные, не игнорируйте их! Худшее, что вы можете сделать, — это предположить, что с ними все в порядке, а затем, спустя несколько недель упорного аналитического труда, обнаружить, что это не так и все усилия были напрасны. В такой ситуации следует поговорить либо со стейкхолдером, либо с кем-то, кто отвечает за эти данные, и спросить, знают ли они о несоответствиях. Во многих случаях об этом уже знают и могут предложить это проигнорировать. В примере с датасетом авиакомпании можно просто удалить данные о бортах, которые приземлились до взлета.

Если выяснится, что о проблеме не знали, а она может поставить под угрозу анализ, необходимо изучить способы спасения ситуации. Если вам нужен срав-

нительный анализ доходов и расходов, но при этом, как ни странно, половина имеющихся у вас данных не содержит информации о тратах, посмотрите, можете ли вы работать только с каким-то одним их видом (например, только с доходами). В некотором смысле этот подход превращается в анализ в квадрате: вы проводите мини-анализ, чтобы понять, возможен ли общий анализ в принципе.

10.3.2. Просмотр и моделирование данных

Во время просмотра и моделирования вы пересматриваете свой план анализа и пытаетесь завершить работу. В следующих разделах представлена общая схема подхода к каждому его пункту.

ИСПОЛЬЗУЙТЕ МЕТОД ОБОБЩЕНИЯ И ПРЕОБРАЗОВАНИЯ

Подавляющее большинство аналитической работы можно выполнить путем обобщения и преобразования данных. Чтобы ответить на вопрос: «Сколько клиентов у нас было каждый месяц?», можно взять информацию о клиентах, сгруппировать ее по месяцам, а затем подсчитать количество человек в каждом месяце. Здесь не нужны статистические методы или модели МО — просто преобразования.

Можно решить, что это не очень-то похоже на анализ данных, ведь вы не пользуетесь ничем, кроме множества арифметических действий, но часто правильное выполнение преобразований бесценно. Большинство других людей в компании вообще не имеют доступа к данным, не могут эффективно их преобразовывать или не знают, какие преобразования вообще нужны.

В зависимости от данных вы можете задействовать некоторые статистические методы, например поиск значений на разных уровнях процентилей или вычисление среднеквадратического отклонения.

ВИЗУАЛИЗИРУЙТЕ ДАННЫЕ ИЛИ СОЗДАЙТЕ СВОДНЫЕ ТАБЛИЦЫ

Выполнив соответствующие преобразования, создайте визуализации или сводные таблицы, чтобы видеть, что происходит с данными. Вернемся к предыдущему примеру: если у вас есть определенное количество клиентов в месяц, можно создать столбчатую диаграмму, чтобы проследить изменения. С помощью такого графика можно легко увидеть закономерности, а не просто вывести фрейм данных на экран.

На рис. 10.4 представлен пример сводной визуализации, показывающей общее количество клиентов за каждый месяц. На этом графике люди могут легко увидеть, что их число понемногу растет.

Выбор типа визуализации зависит от имеющихся данных. Можно использовать линейный график, диаграмму размаха или любой из многих других вариантов.

Вместо диаграммы можно сделать сводную таблицу. Все зависит от того, что вы пытаетесь понять. В конце этого раздела среди материалов к нему можно найти информацию о выборе правильного типа графика для ваших данных. Обратите внимание, что в процессе визуализации вы можете осознать, что некоторые этапы преобразования следует изменить. Скорее всего, вам придется много раз скакать между разными этапами анализа.

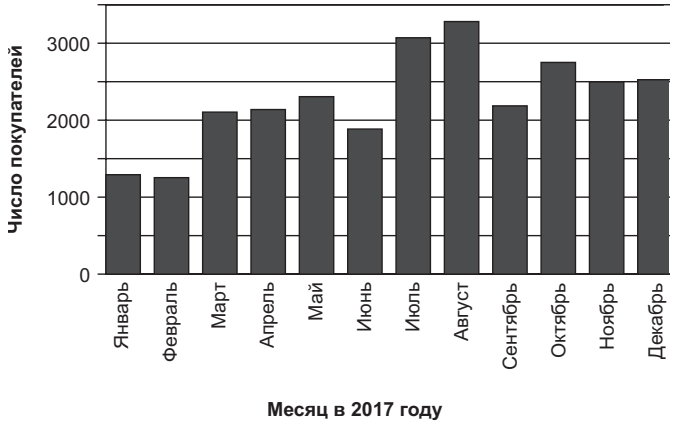


Рис. 10.4. Пример сводной таблицы

Поскольку при визуализации будет множество итераций и непрерывного преобразования данных, следует найти золотую середину между желанием удалить промежуточные этапы ради чистоты кода и стремлением сохранить абсолютно все на всякий случай. Лучше всего оставить как можно больше кода при условии, что (1) старый код будет работать нормально после того, как вы внесете дальнейшие изменения, и (2) вы можете четко обозначить, какие результаты являются «хорошими». Не храните код, не работающий для вашего анализа, или его большие части с комментариями — это чрезвычайно затрудняет его поддержку. Помимо этого, неплохо бы использовать контроль версий вроде git и GitHub; каждый раз, добавляя в анализ новый контент, можно записывать изменения и откатывать код, если что-то пошло не так.

ПРИ НЕОБХОДИМОСТИ СОЗДАЙТЕ МОДЕЛЬ

Если в данных есть закономерности, которые предполагают, что здесь отлично подойдет моделирование, — действуйте! Возможно, было бы целесообразно применить модель временно́го ряда, чтобы спрогнозировать количество клиентов в следующем году. При создании моделей следует выводить результаты и визуализировать их, чтобы понимать, насколько они точны или полезны. Можно

создавать графики, которые сравнивают прогноз с фактическими значениями или содержат такие показатели, как оценки точности и значения важности функций.

Если вы создаете модели машинного обучения, которые могут быть использованы не только для анализа, а, например, для запуска в производство (о чем мы расскажем в главе 11), убедитесь, что вы изолировали код построения модели от общего анализа. Поскольку в дальнейшем вы будете использовать только ее, вам нужно будет легко извлекать этот код из того, который составляет общие диаграммы визуализации.

ПОВТОРИТЕ

Все эти шаги нужно проделать для каждого пункта плана анализа. В процессе работы у вас может появиться новое представление о том, что следует анализировать, или вы вдруг осознаете, что то, что вы считали целесообразным, на самом деле не имеет смысла. В таком случае нужно скорректировать план и продолжить работу.

Вероятно, разные пункты плана анализа будут взаимосвязаны, поэтому код, который вы использовали в одном пункте, будет повторяться и в другом. Постарайтесь структурировать план, чтобы можно было запускать одну и ту же программу несколько раз, а обновления в одной ее части сразу же распространялись бы на другие. Ваша цель — создать систему, которую можно поддерживать и легко менять, не затрачивая массу времени на отслеживание сложного кода.

10.3.3. Важные моменты для анализа и моделирования

Работа по анализу и моделированию данных зависит от поставленной задачи. Математические и статистические методы для кластеризации данных не подходят для прогнозирования или оптимизации решения. Если вы будете следовать некоторым общим советам, то сможете сделать не просто хороший, а отличный анализ.

СОСРЕДОТОЧЬТЕСЬ НА ОТВЕТЕ НА ВОПРОС

Как уже говорилось в разделе 10.2, очень легко потратить время на выполнение работы, которая не приводит к достижению цели. Если вы анализируете заказы клиентов с целью узнать, можно ли спрогнозировать, когда клиент сделает свой последний заказ и больше никогда не вернется, вы можете получить нормально работающую модель нейронной сети, а затем потратить несколько недель на настройку гиперпараметров. Если стейкхолдеру достаточно только ответа «да» или «нет» на вопрос о работоспособности модели, то настройка гиперпараметров для повышения ее эффективности не поможет. Вместо того чтобы неделями заниматься этой настройкой, можно было бы сделать что-то более полезное.

При анализе важно сосредоточиться на плане и ответить на заданный вопрос. Нужно постоянно спрашивать себя: «Это относится к поставленной задаче?» Вы должны думать об этом каждый раз, когда составляете график или таблицу. Здорово, если вы все время положительно отвечаете на этот вопрос. Но ситуации, когда вы порой думаете: «Этот график (или таблица) бесполезен», гораздо более вероятны, и тогда, возможно, придется скорректировать свою работу. Во-первых, постарайтесь остановиться и подойти к проблеме иначе. Если вы пытались сгруппировать клиентов по расходам, попробуйте вместо этого выполнить кластеризацию. С совершенно новым подходом у вас больше шансов на успех, чем при внесении незначительных изменений. Во-вторых, поговорите с руководителем или стейкхолдером проекта: возможно, данные, которые вы используете, неэффективны для решения возникшей проблемы.

На протяжении всего процесса анализа нужно стабильно собирать набор действительно актуальных результатов и (в идеале) следовать плану.

ИСПОЛЬЗУЙТЕ ПРОСТЫЕ МЕТОДЫ ВМЕСТО СЛОЖНЫХ

Сложные методы так увлекательны! Зачем нам линейная регрессия, если можно использовать случайный лес? Зачем использовать случайный лес, если есть нейронная сеть? Эти методы показали себя эффективнее, чем старая добрая регрессия или кластеризация *k*-средних, и они интереснее. Поэтому, когда вас просят решить бизнес-вопросы с помощью данных, вы, безусловно, должны использовать наилучшие методы.

К сожалению, у сложных методов есть множество недостатков, которые неочевидны, если сосредоточиться исключительно на точности. Цель анализа не в том, чтобы получить максимально возможную точность или прогноз; он должен ответить на вопрос так, чтобы представитель бизнеса мог его понять. Это означает, что вам нужно объяснить, откуда взялся такой результат. С помощью простой линейной регрессии можно легко построить графики и понять, как каждая функция повлияла на результат, тогда как при использовании других методов описать причины результатов модели бывает очень сложно, а заказчику будет сложнее поверить в их правильность. Более сложные методы настраиваются дольше, отладка и запуск нейронной сети занимают некоторое время, тогда как линейная регрессия выполняется довольно быстро.

Поэтому при проведении анализа стоит выбирать простые методы как в моделях, так и в преобразованиях и агрегировании как можно чаще. Например, вместо того чтобы отсекать определенный процент выбросов, выполните логарифмическое преобразование или возьмите медианное значение вместо среднего. Если линейная регрессия работает достаточно хорошо, не тратьте время на создание нейронной сети ради незначительного повышения точности. Применение простых

методов там, где это возможно, делает результат намного понятнее для других, а вам будет проще осуществлять поддержку и отладку.

ГРАФИКИ ДЛЯ ИССЛЕДОВАНИЙ VS ГРАФИКИ ДЛЯ ПРЕЗЕНТАЦИЙ

Есть две разные причины, по которым дата-сайентист предпочитает визуализировать данные: для исследования и для обмена информацией. Первый он составляет, чтобы понять динамику данных. Сложный и плохо размеченный граф — это нормально до тех пор, пока специалист понимает его. Цель графика для презентаций состоит в том, чтобы тот, кто мало знает о данных, получил конкретную информацию, которую пытается вывести дата-сайентитст. Для того чтобы такой график считался эффективным, он должен быть простым и понятным. При проведении анализа вам понадобится использовать множество графиков для анализа, но ими не стоит делиться с другими сотрудниками.

Рассмотрим пример, основанный на вымышленных данных о кличках домашних животных в городе: дата-сайентист хочет понять, соотносится ли первая буква клички домашнего животного с его видом (коты или собаки). Он загружает данные и создает визуализацию, показывающую для каждой буквы разбиение групп котов и собак, чьи клички начинаются с нее (рис. 10.5).

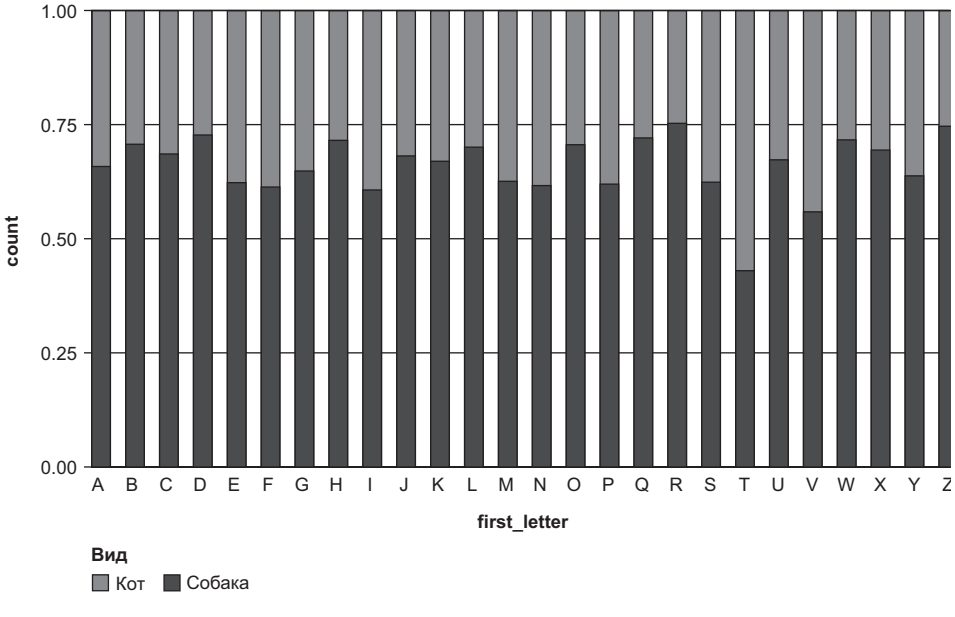


Рис. 10.5. Пример визуализации, сделанной во время анализа перед очисткой

Если вы внимательно посмотрите на рис. 10.5, то заметите, что в столбце T гораздо больше кошек, чем собак — важный вывод для дата-сайентиста. При этом это не тот график, который следует показывать стейкхолдеру: в нем слишком много информации и с первого взгляда не все очевидно.

На рис. 10.6 показаны те же данные, построенные другим, более простым способом. В этой версии очевидно, что шанс котов получить кличку, начинающуюся на букву T, составляет 12 %, тогда как для собак он равен только 5 %. Теперь этими данными можно поделиться со стейкхолдером.

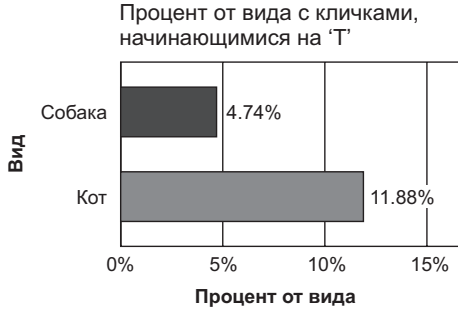


Рис. 10.6. Те же данные, что и на рис. 10.5, составленные так, чтобы подчеркнуть важность буквы T

НЕПРЕРЫВНАЯ ГОТОВНОСТЬ ДЕЛИТЬСЯ

Результат анализа может принимать различные формы, выбор которых зависит от целевой аудитории. Если анализ предназначен для бизнесменов, часто используется набор слайдов или редактируемый документ. PowerPoint или Word (или Google Slides и Google Docs) — хороший вариант, потому что любой может открыть такие файлы (при наличии пакета Microsoft Office для первых двух вариантов), а их форматы поддерживают множество диаграмм, таблиц и текстовых описаний. Если анализ предназначен для технических специалистов, можно передать выходной HTML-файл Jupyter Notebook или R Markdown. Эти методы хороши тем, что для их обработки обычно требуется меньше усилий (например, не нужно тратить время на выравнивание фигур на слайде). Если нужно передать большое количество таблиц с данными для финансовых специалистов, лучшим вариантом может стать Excel: это отличный инструмент на случай, если конечному пользователю нужно взять значения из результатов и произвести дальнейшие вычисления. Чтобы впоследствии вам не пришлось столкнуться с доработкой, решите на раннем этапе анализа, какой результат вы ожидаете получить.

В зависимости от объема работы вам нужно будет периодически связываться с заказчиком и показывать ему результат. Это позволяет предотвратить катастрофу:

представьте, что вы неделями в одиночку работали над анализом, а когда пришло время сдавать проект, стейкхолдер находит что-нибудь, что сводит все ваши усилия на нет. Например, говорит: «Вы рассматривали продажи, но не учли возвраты». Если бы этот пункт был указан в начале, напрасной траты времени можно было бы избежать. Кроме того, стейкхолдер часто может вносить изменения, предлагая возможные сферы, на которых необходимо сосредоточиться, или методы, которые следует попробовать. В некотором смысле такое общение на протяжении всего процесса аналогично концепции гибкой методологии разработки ПО: постоянное улучшение работы, а не выпуск одного объемного готового продукта.

Регулярная координации со стейкхолдером — это хорошо, но дата-сайентисты часто пренебрегают им. А еще в этой ситуации плохо то, что работу приходится показывать человеку, не занимающемуся данными, то есть промежуточный результат должен быть в таком состоянии, которое не стыдно показать. Нужны графики с четкими обозначениями и смыслом, код с минимальным количеством ошибок и базовая история происходящего. Поэтому дата-сайентисты часто думают: «Не буду показывать результат, пока не доработаю его; займусь этим в конце». Не делайте так! В итоге это практически всегда выливается в больший объем работы. С постоянным поддержанием работы в виде, подходящем для показа, вы в итоге получите более качественный продукт.

ЗАПУСК ОДНОЙ КНОПКОЙ

Точно так же как для загрузки и подготовки данных неплохо бы запустить только один сценарий, нужно предусмотреть единственную кнопку, запускающую анализ. В Python для автоматической загрузки данных есть Jupyter Notebook, который выполняет анализ без ошибок. В R это файл R Markdown — он загружает данные, анализирует их и выводит HTML-файл, документ Word или презентацию PowerPoint.

При проведении анализа избегайте запуска слишком большого количества кода за пределами сценария или беспорядочного выполнения сценариев. Такие методы увеличивают вероятность ошибки при повторном запуске. Это нормально, если вы захотите написать ad hoc код под конкретную ситуацию, но для начала убедитесь, что сможете повторно запустить файл без ошибок. Так вы сможете в любой момент поделиться результатами работы с другими людьми и гарантированно потратите меньше времени на исправление сценария в конце анализа.

10.4. Завершение

В зависимости от требований стейкхолдера результатов выдачи кода может оказаться достаточно для удовлетворения запроса, но, возможно, вам придется пойти дальше и подготовить окончательную версию. Если финальный результат должен

быть представлен в виде презентации PowerPoint, скорее всего, надо будет его отшлифовать чуть больше, чем при проведении анализа, чтобы он соответствовал требованиям компании к стилю. Самое главное, вам нужно будет составить рассказ для окончательного документа, чтобы люди, не участвовавшие в процессе, могли полностью понять ваши выводы, а также что и для чего было сделано.

Такой рассказ — первый шаг к получению хорошего итогового документа. Что вы собираетесь сообщить? Как вы представите проблему, как объясните решение, которое предлагает проделанная работа (или не предлагает), и как обсудите следующие шаги? Рассказать о работе можно по-разному, но один из самых простых способов — подумать, как можно объяснить процесс человеку, который раньше с ним не сталкивался. Подумайте, что бы вы рассказали, и постарайтесь сделать то же самое в своем документе. Неоднократно задавайте себе эти вопросы: «Будет ли то, что я показываю, понятно моей аудитории?» и «Что я могу сделать, чтобы улучшить рассказ?» В конце концов вы достигнете того момента, когда будете довольны контентом.

Вам также нужно будет добавить в свой документ текст. Обычно это делается, чтобы прокомментировать ваш рассказ или пояснить, почему важна каждая диаграмма. Опять же, постарайтесь сделать его понятным для человека вне контекста. Спросите себя: «Чем моя информация полезна для бизнеса?» Требования к количеству текста различаются между компаниями: некоторые хотят подробных описаний, объясняющих все, в то время как другим достаточно нескольких слов. Лучше написать побольше, потому что впоследствии можно сократить содержание.

Когда основной материал готов, проверьте работу на предмет мелких ошибок, прежде чем отправлять его стейкхолдеру. Попросите заняться этим коллегу, знакомого с контекстом работы, чтобы уточнить, все ли в порядке. В некоторых компаниях руководитель может потребовать, чтобы для начала работу предоставили ему на проверку для утверждения.

10.4.1. Итоговая презентация

Когда ваш руководитель утвердит анализ, назначьте встречу со стейкхолдером, чтобы представить работу лично. Подробно расскажите ему о каждом компоненте и опишите, что вы сделали, что узнали и на что решили не обращать внимания. Вы потратили так много времени на анализ данных, что рассказать о нем и ответить на пару вопросов не должно быть трудно.

В зависимости от стейкхолдера вас могут засыпать вопросами на протяжении всей презентации или же человек может ждать, пока вы закончите. Они могут быть как спокойными и выражать любопытство («Почему вы использовали набор данных X вместо набора данных Y?»), так и содержать критику и озабоченность («Почему у вас не такие результаты, как у другой команды? Есть ли в вашем коде

ошибки?»). Ответы во многом аналогичны ответам во время интервью (глава 7): вы можете честно сказать о том, что знаете и что вам неизвестно. Можно упомянуть, что вам нужно что-то дополнительно изучить. Будьте максимально открыты в обоснованиях («Мы использовали набор данных X, потому что он охватывает интересующий нас период») и скажите честно, если чего-то не знаете («Я не уверен, почему эти выводы не соответствуют результату, полученному другой командой. Я займусь этим вопросом»). При этом в большинстве случаев такие встречи проходят спокойно и бесконфликтно!

Независимо от того, насколько хорош анализ, вас обязательно спросят: «А как насчет _____?», где пробел — это то, на что вы не обращали внимания в процессе. Кто-то может спросить: «А как насчет того, чтобы анализировать только данные за последний месяц?» В Data Science это естественно: всегда можно выделить какую-нибудь полезную вещь. Особенно часто это встречается в ситуациях, когда анализ оказался безрезультатным. В таких случаях человек, обращающийся с вопросом, всеми силами пытается найти зацепку, которая может оказаться решающей.

Как дата-сайентист лучшее, что вы можете сделать в таких ситуациях, — это попытаться мягко противостоять подобным вопросам. Несмотря на то что в некоторых случаях они оказываются полезными, не исключено, что вы снова окажетесь без каких-либо новых выводов и только потеряете время. Как специалист вы должны лучше других знать, что может подойти, и, если вы не считаете, что предложенный вариант пойдет на пользу, скорее всего, так оно и есть. Вопрос, который вы пытаетесь решить, часто бывает настолько абстрактным, что невозможно дать по-настоящему однозначный ответ. Завершив анализ, вы должны уметь вовремя остановиться, равно как и в случае с выбором методов анализа, когда не следует перебирать все подряд.

10.4.2. Длительное хранение работы

Когда окончательный анализ передан и одобрен, вам тут же предложат взяться за новую задачу. Однако прежде, чем приступать к ней, сделайте несколько маленьких вещей, чтобы в будущем упростить себе жизнь. Велика вероятность, что через несколько месяцев или даже лет вас попросят повторить анализ, но уже с более свежими данными. Если вы сохраните свою работу, то выполнить задачу будет намного проще. Итак:

- *Еще раз проверьте, можно ли повторно запустить весь анализ*. Ранее мы говорили, что следует запускать анализ одной кнопкой; на этом этапе вы должны сделать последнюю проверку и убедиться, что все работает.
- *Прокомментируйте свой код*. Поскольку вы можете забросить этот код в дальний угол на несколько лет, даже небольшие комментарии могут помочь вспомнить, как использовать или изменять его.

- *Добавьте файл README*. Файл README — это простой текстовый документ, в котором описывается, для чего нужен анализ, почему он был проведен и как его запустить.
- *Храните код в надежном месте*. Считайте, что вам это уже удалось, если вы используете git и GitHub, но если нет, подумайте, как можно будет получить доступ к коду через долгое время.
- *Убедитесь, что данные надежно сохранены*. Проверьте, что все файлы данных хранятся в надежном месте, и это, естественно, не ваш ноутбук, а облачные сервисы (например, OneDrive, общий сетевой диск или AWS S3). Кроме того, в идеале следует проверять датасеты из БД, чтобы убедиться, что никто их не удалит.
- *Результат хранится в общедоступном месте*. Чаще всего люди обмениваются результатами анализа в виде вложений в имейлах, но это не лучший способ архивации. Разместите результаты в месте, к которому могут получить доступ другие члены команды и люди из других подразделений компании.

Когда эта работа будет сделана, можете считать, что анализ действительно завершен. По мере роста количества анализов вы найдете методы и способы, которые подходят вам лучше всего, что, несомненно, ускорит и улучшит вашу работу.

10.5. Интервью с Хилари Паркер, дата-сайентистом в Stitch Fix

Хилари Паркер (Hilary Parker) работает в Stitch Fix, онлайн-сервисе по персональному стайлингу, и занимается созданием моделей машинного обучения, предлагающим покупателям подходящую одежду. До этого она была старшим аналитиком данных в Etsy. Хилари Паркер получила степень кандидата наук в области биостатистики в Блумбергской школе общественного здравоохранения Университета Джонса Хопкинса.

Как забота о задачах других людей сказывается на анализе?

Практически каждый анализ я начинаю с попытки понять, «кто чего хочет». Например, не потому ли появилась эта задача, что продакт-менеджеру нужно принять решение, а он не готов это сделать, пока не получит анализ эксперимента? Если мы хотим реализовать стратегическую концепцию, то нужно ли доказать, что это принесет X долларов за Y лет, чтобы ее приняли? Я обязательно сажусь и говорю с потенциальными потребителями анализа, чтобы разобраться в ситуации.

Когда вы представляете продукт, крайне важно понимать аудиторию, ее состояние и цели. Хотят ли слушатели углубляться в скучную конкретику? Что бы им понравилось больше? Если вам кажется, что нужно дать больше информации,

можете предложить более подробные статистические данные, но если интерес начинает пропадать, вернитесь к менее подробному описанию.

Как вы структурируете свой анализ?

Я считаю, что важно структурировать анализ так, чтобы он был доступным. Стараюсь дать краткую выдержку в начале и не использую сложные графики, потому что большинство людей не могут быстро их усвоить. А еще я не составляю заметки в виде потока сознания, а это, как я погляжу, делают многие. Записи выглядят как текст, так что они просто добавляют все больше и больше разъясняющих комментариев и выдают все это как результат. В итоге получается что-то вроде: «Вот с чего я начал, и вот что у меня получилось». Но на самом деле это нужно перевернуть так: «Вот заключение, а в приложении вы увидите, с чего я начал». Имейте в виду, что это будет читать живой человек и то, что будет быстрее и проще для вас, для других может оказаться недостаточно понятным. Я настолько сосредоточена на окончательном формате, что это становится частью рабочего процесса. Мне никогда не приходится наводить красоту в блокноте с сотней непонятных заметок: я сразу пишу красиво.

Как вы «шлифуете» итоговую версию?

Я считаю, что цвета имеют действительно большое значение. У многих компаний есть корпоративная тема, например Stitch Fix использует свои фирменные цвета. У нас есть шаблоны ggplot2, в которые встроены цвета из нашей палитры. Подобные вещи действительно эффективны, потому что позволяют людям чувствовать свою причастность к компании. То же самое и с презентациями Google Slides: там есть шаблоны, которые чаще используют из-за их удачного оформления.

Я также придерживаюсь принципа «не переусердствуйте». Одним из моих первых проектов в Stitch Fix было открытие направления одежды больших размеров. Нам нужно было провести экспресс-анализ, чтобы понять, правильные ли размеры мы выбрали. У меня ушло много времени на создание своей маленькой системы по проведению анализа. Я была в восторге от разработки этого воспроизводимого веб-сайта, который будет динамически обновляться каждые X часов, чтобы показать изменения. Но в конечном итоге люди, с которыми я работала, не придали этому большого значения. Я была поглощена созданием веб-сайта, вместо того чтобы поддерживать контакт с партнерами. С эстетикой анализа легко переборщить. Делайте столько, сколько требуется, и не более.

Как вы работаете с людьми, которые просят внести коррективы в анализ?

В последнее время я много читала о дизайн-мышлении, и в контексте проектирования это происходит постоянно. Описанная там точка зрения совпадает с моей собственной: люди плохо умеют излагать свои мысли, не пытаются посмотреть

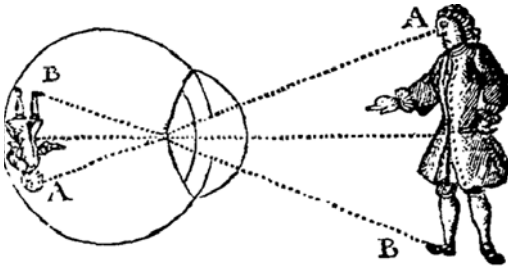
на проблему со стороны и думать абстрактно. В мире дизайна человек приходит к вам и говорит, что ему нужно, но не стоит воспринимать его слова буквально. Вы должны помочь ему сформулировать задачу. В частности, в этом состоит ценность проектировщиков: они рассматривают задачу целостно и систематически перебирают различные варианты решения, пока не найдут наиболее рациональный.

Думаю, это также применимо к дата-сайентистам и специалистам по статистике. Кто-то просит вас добавить в анализ какой-то параметр, потому что человека беспокоит какой-то вопрос, а эта просьба — один из способов выражения беспокойства, но вы должны понять, в чем истинная проблема. Вы хотите сказать, что не готовы принять решение? Сомневаетесь? Какой конечный результат вам нужен? Как специалист по данным вы практически всегда взаимодействуете с одним из потребителей. Вы должны не просто делать то, что вам говорят, а выяснять, что люди пытаются сказать на самом деле. В чем первопричина их слов? Подходит ли это для анализа? Вариантов может быть множество, и вместо того, чтобы просто перебирать их до бесконечности, бывает очень полезно взглянуть на ситуацию с высоты птичьего полета.

Итоги

- Анализ — это документ, в котором представлены выводы и важные особенности применения Data Science для решения задач бизнеса. Он крайне важен для специалистов по данным.
- Хороший анализ требует понимания бизнес-вопроса и того, как на него ответить с помощью данных.
- При проведении анализа всегда думайте о конечной цели, используйте простые методы с четкой визуализацией и будьте готовы поделиться своей работой.
- Важно управлять аналитическим процессом для того, чтобы работа двигалась к конкретной цели и имела логичное завершение.

11



Развертывание модели в производство

В этой главе

- Построение модели машинного обучения для использования в производственной среде.
- API-интерфейсы и их польза.
- Развертывание модели машинного обучения.

Эта глава стремится охватить основные концепции работы инженера по машинному обучению — человека, который создает модели МО и развертывает их для использования в бизнесе. Если ваша работа предполагает создание анализов и отчетов, не пугайтесь! Разрыв между специалистом по принятию решений и инженером МО меньше, чем кажется, и эта глава станет полезным введением в такие вещи.

Иногда цель проекта по анализу не в том, чтобы ответить на вопрос с помощью данных, а в создании инструмента, который использует модель машинного обучения для чего-то полезного. С одной стороны, можно провести анализ, чтобы понять, какие товары люди склонны покупать вместе, а с другой — создать программу, которая рекомендует покупателю наиболее подходящий вариант на веб-сайте. Работа по созданию модели МО, которую можно будет использовать в том числе и в других частях бизнеса, например на веб-сайте или в колл-центре, обычно сложна. Над ней работают дата-сайентисты, инженеры ПО и продакт-менеджеры.

В этой главе мы обсудим, как проектировать модели в виде части продукта и как перенести их с вашего ноутбука туда, где они смогут работать.

Две небольшие заметки перед тем, как мы погрузимся в эту тему:

- Поскольку задача написания кода, работающего в производственной среде, носит довольно технический характер, в этой главе тоже больше технических моментов по сравнению с другими. Поскольку мы не хотим усложнять жизнь людям, которые плохо знакомы с понятиями разработки ПО, то сосредоточимся больше на концепциях и идеях, чем на технических деталях.
- Раз мы уделяем больше внимания понятиям, иногда мы делаем общие высказывания, которые могут не на 100 % соответствовать действительности. Это сделано для удобства восприятия текста. Если вы уже знакомы с этими темами и можете доказать обратное в противовес тому, что мы написали, вы наверняка правы!

11.1. А что вообще развертывается в производство?

Когда говорят «развернуть в производство», имеется в виду взять код и поместить его в какую-то систему, которая позволяет ему работать непрерывно, обычно как часть продукта, ориентированного на клиента. Развертывание — это глагол, это акт перевода его в другую систему. А производство — существительное, это место, где выполняется код, являющийся частью продукта. Код, который находится в производстве, должен иметь возможность работать с минимальным количеством ошибок или проблем, потому что клиенты могут заметить сбои.

Хотя разработчики ПО запускали код в производство на протяжении десятилетий, для дата-сайентистов, в частности инженеров МО, становится все более привычным тренировать модели машинного обучения и запускать их в производство. Обучение модели для запуска в производство аналогично обучению модели в рамках анализа, но в первом случае в дальнейшем требуется пройти значительно больше этапов, чтобы подготовить ее к производству. Часто процесс построения модели для производства начинается с анализа. Во-первых, вам нужно понять данные и заручиться поддержкой бизнеса, а уж позже можно подумать о развертывании в производственной среде. Таким образом, эти два действия очень тесно взаимосвязаны.

Приведем краткий пример, чтобы вы могли лучше понять, что означает развертывание в производство. Предположим, будто некий специалист считает, что компания теряет слишком много покупателей, и поэтому он просит дата-сайентиста провести анализ оттока клиентов. Тот строит модель и показывает, что существует несколько ключевых метрик оттока. Заказчику понравился анализ, и он понимает, что если бы специалисты по обслуживанию клиентов, работающие в колл-центре, знали, какие покупатели, скорее всего, уйдут, они могли бы предложить скидки ради их удержания.

На этом этапе дата-сайентисту необходимо запустить модель в производство. До этого она хранилась у него на ноутбуке, а теперь должна каким-то образом запускаться и вычислять вероятность оттока каждый раз, когда клиент обращается в службу поддержки. На ноутбуке модель обрабатывала множество человек за пару минут, но в производстве придется работать с единственным покупателем во время звонка, извлекая его данные из других частей компании и используя их для составления рейтинга.

Большинство продукционных моделей машинного обучения похожи: они должны работать практически в реальном времени, чтобы делать прогнозы или классифицировать что-либо на основе предоставленных данных. К известным примерам можно отнести модель рекомендаций фильмов Netflix, которая прогнозирует, какие фильмы хотел бы смотреть человек, модель распознавания лиц Facebook, которая берет изображение, находит на нем лица и сопоставляет их с пользователями, и модель автозаполнения Google Gmail, которая анализирует текст по мере ввода и прогнозирует следующее слово.

Продукционные модели должны пройти несколько существенных этапов. Во-первых, они должны быть запрограммированы на обработку любых возможных сценариев во время выполнения кода, чтобы быть менее подверженным ошибкам. При проведении анализа небольшие объемы неадекватных данных можно отфильтровать и не вводить в модель, чтобы не нарушать результат. В продукционной модели код должен выполняться независимо от неадекватности входных данных. В анализе считается нормальным, если модель обработки с использованием естественного языка дает сбой при запуске кода со смайликами, потому что вы можете просто игнорировать такие элементы. А вот если в продукционной модели код дает сбой из-за смайликов, это может привести к нарушению в продукте, который она поддерживает. Представьте, что произойдет, если веб-страница Gmail будет выходить из строя каждый раз, когда вы вводите смайлик. Продукционные модели должны быть разработаны так, чтобы они сами обрабатывали несоответствия, или же код должен их исправлять, прежде чем они в принципе туда попадут.

А еще продукционные модели должны быть удобными в сопровождении. Поскольку они постоянно используются в продуктах, их периодически приходится переучивать на обновленных данных или кодировать так, чтобы это происходило автоматически. Необходимы способы контроля качества работы моделей, чтобы сотрудники компании могли понять, что какой-то процесс нарушен или вообще перестал выполняться. И поскольку они, возможно, понадобятся в течение многих лет, их код должен быть написан так, чтобы эти модели соответствовали стандартам остальных и могли обновляться. Создание кода на мертвом языке, который мало кто знает, плохо для анализа и катастрофично для продукционной модели. Посмотрите на рис. 11.1, где показан процесс создания и развертывания продукционной модели машинного обучения.

Рис. 11.1. Пример создания продукционной модели машинного обучения

В оставшейся части этой главы рассматриваются три концепции: как создать модель машинного обучения, подходящую для производственной среды, как развернуть ее в производство и как поддерживать ее работоспособность.

Производство для разных типов дата-сайентистов

В зависимости от типа специалиста по данным, к которому вы относитесь, вы будете взаимодействовать с продукционными системами следующим образом:

- *Инженер по машинному обучению*. Продукционные системы — это практически вся ваша работа. К тому времени, как вы освоитесь в своей роли, вы должны будете разобраться во всем, что написано в этой главе.
- *Аналитик*. Аналитику, возможно, придется иметь дело с продукционными системами в зависимости от схемы отчетности. Если аналитическая группа регулярно пишет отчеты, лучше всего создать для них системы. Обеспечив автоматическое обновление отчетов, аналитики могут выполнять другую работу. Это особенно актуально для сводных панелей, когда предполагается, что системы находятся в производстве и обновляются самостоятельно.
- *Специалист, принимающий решения*. Поскольку такой сотрудник в основном занимается решением ad hoc случаев, у него не так много возможностей для создания продукционных систем. Но если их модели будут переданы инженерам МО для запуска в производство, то неплохо бы понимать их получше. Специалисты по принятию решений также могут создавать интерактивные инструменты для бизнеса из библиотек вроде Shiny или Dash, которые необходимо развертывать и поддерживать в продукционных системах.

11.2. Создание продукционной системы

Создание продукционной системы, основанной на модели МО, начинается с тех же шагов, что и в случае с моделями для анализа: нужно найти подходящие данные, выделить признаки, обучить модель и обеспечить вовлеченность бизнеса. Но в конце потребуется кое-что еще:

1. Модель необходимо конвертировать в формат, подходящий для других программ. Как правило, для этого пишется код, позволяющий обращаться к модели как к API из других систем компании, как если бы это был вебсайт.

2. В модель должен быть добавлен код для обработки множества потенциальных входных данных. Это гарантирует, что неожиданные значения не приведут к сбою и значительно сократят время простоя. Для этого в модель должно быть добавлено тестирование, обеспечивающее проверку правильности обработки всех данных.
3. Модель развертывается в тестовой среде для проверки правильности ее работы. Проводится тестирование API, чтобы убедиться, что он функционирует и может обрабатывать объем трафика, который будет поступать в реальной ситуации.

Когда все эти этапы пройдены, модель развертывают в производственной среде.

11.2.1. Сбор данных

Когда вы собираете данные для обучения модели при анализе, вам необходимо найти подходящий архивный датасет с хорошим сигналом. То же самое необходимо и для продукционной модели, но часто этого недостаточно, потому что необходимо учитывать ее компонент, работающий в реальном времени. Рассмотрим предыдущий пример, где компании нужна модель в производстве, которая в реальном времени прогнозирует уход клиента. Если бы она была нужна только для анализа, то архивный набор атрибутов покупателя (количество покупок, количество времени с момента первой покупки и так далее), полученный несколько месяцев назад, отлично подошел бы. Но поскольку продукционная модель должна делать прогноз в реальном времени, ее код необходимо написать так, чтобы каким-то образом заполучить атрибуты покупателя в момент вызова. Если клиент #25194 звонит в службу поддержки, код должен получить точные данные о его заказах, чтобы модель смогла осуществить прогноз.

Между использованием архивных данных для обучения модели и предоставлением информации в режиме реального времени при ее запуске разница может быть огромной. По техническим причинам, связанным со способом сбора данных, может пройти несколько часов или даже дней, прежде чем они поступят в БД или другое хранилище, к которому будет доступ у соответствующего специалиста. Кроме того, бывают ситуации, когда данные доступны в реальном времени, но не записываются в архиве. Например, вы можете узнать, находится ли клиент сейчас за рубежом, но информация о его прежних выездах из страны не сохранилась.

Когда вы ищете данные для создания продукционной модели, подумайте, что потребуется в реальном времени при запуске. Будет ли информация достаточно актуальной? Можно ли получить к ней доступ через соединение с БД или каким-либо другим способом и кто его может настроить? Проекты машинного обучения нередко проваливаются из-за проблем с датасетами.

11.2.2. Построение модели

Если у вас есть подходящий набор данных, можно приступать к построению модели машинного обучения. Это обширная тема; если вы хотите узнать, как построить модель МО от разработки функций до обучения и проверки, для этого есть множество книг и интернет-ресурсов. Вместе с тем при создании модели специально для производства нужно учитывать несколько вещей.

УДЕЛИТЕ ОСОБОЕ ВНИМАНИЕ КАЧЕСТВУ МОДЕЛИ

Поскольку другие системы будут зависеть от того, насколько эффективно ваша модель обрабатывает любые поступающие из них данные, нужно обязательно понимать принципы ее работы. Допустим, вы создаете модель МО как часть анализа, чтобы понять, какие продукты интересуют клиентов: например, вам нужно спрогнозировать, какой продукт покупатель приобретет в следующий раз. Если ваша модель правильно предсказывала 99 % покупок, но в 1 % случаев сообщала, что человек закажет временные татуировки с Николасом Кейджем, она будет иметь огромный успех. Понимая потребности большинства покупателей, вы можете помочь компании принять обоснованное маркетинговое решение. Но если вы собирались развернуть эту модель, чтобы показывать рекомендуемые продукты на веб-сайте компании, этот 1 % может быть катастрофическим и привести к потере клиентов (или по крайней мере тех из них, которые не ценят непревзойденную харизму Николаса Кейджа). То, что происходит на периферии, действительно важно в продукционных системах, но не в анализе.

ПОСТРОЕНИЕ ПРОСТОЙ МОДЕЛИ

Как только модель будет запущена в производство и клиенты начнут с ней взаимодействовать, вы неизбежно столкнетесь с ситуацией, когда с ней происходит что-то странное. Ваша задача — понять почему. Если это модель оттока клиентов, она может предсказывать, что каждый покупатель на Аляске откажется от услуг по неизвестной причине. Или же система рекомендаций на веб-сайте товаров для активного отдыха может предлагать только каяки. В этом случае вам нужно будет копнуть глубже и выяснить, что происходит, а также нужно ли каким-либо образом изменить модель.

При использовании простой модели, например на основе линейной регрессии, проследить вычисления для прогноза должно быть довольно просто. А вот со сложными методами, например статистическим ансамблем или бустингом, понять происходящее намного сложнее. Несмотря на то что вам следует использовать достаточно сложную модель для решения поставленной задачи, она все же должна быть максимально простой из возможных, даже если это будет стоить точности.

Одна интересная история из реального мира связана с Netflix Prize. Компания Netflix организовала конкурс, чтобы узнать, сможет ли команда разработчиков создать алгоритм, который улучшит результаты рекомендаций фильмов на 10 %, и в 2009 году присудила за это приз в размере $1 000 000. Как сказано в статье Wired (https://www.wired.com/2012/04/netflix-prize-costs), Netflix так и не воспользовалась победившим алгоритмом. Он представлял собой ансамблевый метод, объединяющий множество моделей, но сложность запуска и отладки была настолько высокой, что повышение точности того не стоило. Несмотря на то что в Netflix заплатили огромные деньги за точную модель, компания поняла, что есть вещи поважнее точности. Если даже Netflix со своей армией специалистов по данным и инженеров не поддерживает чересчур сложные решения, то вряд ли мы можем советовать это другим компаниям.

11.2.3. Обслуживание моделей с API

На момент написания этой книги большинство моделей машинного обучения обслуживались как интерфейсы прикладного программирования (application programming interfaces, APIs). Это значит, что их код может работать в одной компьютерной системе и что другие системы при необходимости могут к нему подключаться, чтобы модель работала на их данных. Предположим, у компании есть система, управляющая веб-сайтом магазина, и руководство хочет добавить модель МО, предлагающую скидки клиентам, собирающимся отказаться от подписки. Вместо того чтобы вписывать ее в код веб-сайта, можно создать вторую систему, содержащую модель, с которой этот веб-сайт будет периодически связываться.

Концепция разбиения различных частей системы на небольшие микрослужбы широко используется в разработке ПО и подробно описана во многих книгах. Для дата-сайентистов важна концепция создания одной компьютерной системы, которая работает исключительно с моделью, для того чтобы другие системы могли ее использовать.

Современные API создаются с помощью веб-служб, которые в просторечии называются REST API. REST API — это такой крошечный веб-сайт, но, в отличие от него, он возвращает не HTML для рендеринга в браузере, а данные, обычно в виде форматированного текста. Эти запросы используют протокол HTTP, аналогичный для веб-браузеров (и именно поэтому адреса веб-сайтов начинаются с http:// или https://). Например, API сайта погоды можно настроить таким образом, чтобы при переходе по URL http://exampleweather.com/seattle_temperature показывалась температура в Сиэтле в градусах (45). Для API модели МО нужно перейти на определенный веб-сайт и получить прогноз. В случае с моделью МО где-то в сети компании может быть сайт, который предсказывает, уйдет покупатель или нет. Такой сайт, как http://internalcompany.com/predict?customer=1234, вернет число от 0 до 1, представляющее вероятность ухода этого клиента.

Разработка API включает принятие решений: например, какие URL-адреса будут возвращать данные (и в каком виде) или какие типы запросов следует использовать. Понятный дизайн — важная часть на пути к тому, чтобы люди действительно пользовались им, и зачастую продумывание интерфейса требует не меньшего внимания, чем создание самой модели.

Запуск модели МО как веб-службы API удобен по нескольким причинам:

- Поскольку это веб-API, модель можно использовать где угодно и кем угодно, в том числе другими продуктивными системами или дата-сайентистами. Запрос к той же модели прогнозирования оттока, которая используется на веб-сайте, может быть отправлен специалистом по принятию решений.
- Поскольку модель работает как веб-сайт, ее можно подключить практически к любой современной технологии, независимо от технической платформы. Если модель написана на R, ее по-прежнему может использовать как веб-сайт, написанный на Node.js, так и аналитик, работающий с Python. Кроме того, если по какой-либо причине модель перестанет работать, она с меньшей вероятностью потянет за собой другие продукты, поскольку размещается на собственном веб-сайте. Если на веб-сайте интернет-магазина компании внезапно пропадает связь с моделью, то он все равно должен работать.

11.2.4. Построение API

API-интерфейсы отлично подходят для моделей МО, но для них требуется дополнительный код. И у R, и у Python есть пакеты — Plumber и Flask соответственно, которые делают его за вас. Когда вы запускаете сценарий R или Python с этими пакетами, он направляет функцию в конечную точку на компьютере. Можно указать, что переход по URL-адресу http://yourwebsite.com/predict запустит функцию машинного обучения R или Python, а затем вернет любой ее результат. Затем можно войти в браузер и вызвать свой код! Предположим, вы запускаете его на ноутбуке или ПК. Если вы настроите брандмауэр и предоставите доступ для внешнего подключения к компьютеру, другие люди смогут запустить ваш API. Но, как только вы перестанете запускать программу хостинга API (R или Python), никто не сможет запустить вашу модель.

Хотя и R, и Python упрощают обслуживание API модели, необходимо принимать проектные решения, например какую информацию необходимо будет передать в качестве входных данных в модель в запросе API. Предположим, вы создаете модель, которая предсказывала бы вероятность ухода клиента. Прогноз должен строиться на основании длительности отношений покупателя и компании, потраченной суммы на товары и количества звонков в службу поддержки. Один из возможных вариантов дизайна API — это сделать запрос с уникальным идентификатором клиента в URL. Например, чтобы найти кли-

ента с идентификатором 1234, можно перейти по ссылке http://yourwebsite.com/predict?customer_id=1234. Другой вариант — пользователи должны сами искать всю информацию о клиентах, а затем включать эту информацию в текст запроса. Таким образом, для клиента, пробывшего в компании 1,7 года, с общими затратами в $1257 и тремя звонками в контакт-центр вы можете отправить запрос на http://yourwebsite.com/ и предсказать, где будет тело запроса `{"tenure":1.7,"spend":1257,"calls":3}`.

Оба варианта подходят для дизайна API, но один требует, чтобы API выполнял всю работу по поиску сведений о клиенте, а другой заставляет пользователя API искать информацию. Как правило, принимать подобные решения в одиночку — не лучшая идея; чем больше людей, которые могут использовать ваш API, вы можете вовлечь, тем больше шансов, что они будут довольны результатом.

После того как вы создали API, поговорите с его будущими пользователями и покажите им, как он работает. У них должна быть возможность дать отзывы о дизайне. В идеале вы также должны поделиться с ними документацией по API.

О plumber — пакете R для обслуживания кода в качестве веб-API (Джефф Аллен)

Джефф Аллен (Jeff Allen) работает в RStudio и является создателем plumber — пакета, который позволяет создавать веб-API в R, чтобы модели МО можно было использовать в масштабах всей организации.

Изначально в создании plumber участвовало совсем немного инженеров, зато в него не было вложено много средств; у него было более скромное происхождение. В 2012 году я работал в группе биостатистики в исследовательском центре. Эта группа использовала R для большей части анализа и пригласила меня помочь им в создании улучшенного ПО. Поначалу у нас были три разные аудитории:

- Другие специалисты биостатистики, которые использовали R и хотели применять разработанные нами методы.
- Пользователи, далекие от технической сферы, например врачи, которым просто нужны были результаты анализа, чтобы ответить на вопросы типа: «Какой препарат будет наиболее эффективен для этого пациента?»
- Технические пользователи, которым просто нужен был анализ, при этом они не интересовались или не имели достаточно вычислительных ресурсов для запуска кода R.

Первая аудитория была самой легкой; R поставляется с надежной системой, которая позволяет объединять код и данные в пакеты: ими можно поделиться или передать другим людям для работы. В то время обслужить вторую аудиторию было немного сложнее. Сегодня Shiny является очевидным решением этой проблемы и предлагает пользователям R удобный способ продолжить работу в R для создания многофункциональных интерактивных веб-приложений, которые подойдут для нетехнических специалистов.

Учесть потребности третьей аудитории было сложнее всего. У некоторых пользователей уже было существующее приложение, написанное на другом языке, например на Java, но они хотели вызвать некоторые функции R из своей службы. У других был простой автоматизированный конвейер, и они хотели использовать функцию R с интенсивными вычислениями, которую мы определили. Во всех этих случаях пользователи хотели чего-то такого, что можно было бы вызывать удаленно, и при этом рассчитывали на нас, чтобы мы изнутри выполняли всю обработку в R, а затем отправляли им результат. Короче говоря, им был нужен удаленный API для R.

Спустя годы у меня действительно появилась возможность начать работу над пакетом, который превратится в plumber. Здесь у меня был личный интерес. В большинстве организаций есть группа людей, которые не знакомы с R, но им может пригодиться анализ, который дата-сайентисты создают на этом языке. Для многих требуется программный и структурированный интерфейс, а веб-API предлагают элегантное решение. К счастью, авторы пакета Shiny уже решили все сложности, связанные с созданием высокопроизводительного веб-сервера, который мог бы использоваться пакетами R для обслуживания HTTP-запросов. Оставалось создать интерфейс, с помощью которого пользователи могли бы определять структуру и поведение своего API.

Я надеюсь, что plumber предложит решение для этой технической аудитории, чтобы ее представители могли пользоваться R так же эффективно, как и прочие вышеупомянутые группы. Наблюдая за развитием plumber на протяжении многих лет как с точки зрения функций, так и с точки зрения использования, я думаю, что он понравился пользователям R. Поскольку R может быть удобно реализован через API, теперь он получил место наряду с другими языками, более знакомыми традиционным IT-организациям. Было забавно наблюдать, как люди используют plumber для выполнения задач, с которыми я никогда бы не справился.

11.2.5. Документация

Когда у вас есть работающий API, самое время написать для него документацию. Чем раньше вы это сделаете, тем проще будет в дальнейшем поддерживать API. На самом деле это отличный способ создать документацию, прежде чем писать первую строку кода. В этом случае документация — это ваш план по созданию API, и у людей, которые будут использовать вашу модель, будет достаточно времени, чтобы подготовиться и все изучить.

Основой документации API является спецификация его запросов: какие данные можно отправлять, на какие конечные точки их можно отправить и какого ответа следует ожидать? Эта документация позволяет другим людям писать код, который будет вызывать API, и знать, чего ожидать. Она должна включать множество деталей, например:

- URL-адреса конечных точек: (http://www.companywebsite.com/example);
- что нужно включить в запрос;
- формат и содержание ответа.

Эта документация может находиться в любом текстовом документе, но есть также стандартные шаблоны для ее хранения, например документы OpenAPI. *Документ OpenAPI* — это спецификация для написания файлов спецификации API, которые будут понятны пользователями или компьютерным системам.

В идеале вы не будете постоянно поддерживать работу API, поэтому вам понадобится документация о том, какие требования API предъявляет к системе и как его установить в другом месте. Благодаря такой документации человек сможет самостоятельно запустить код и вносить изменения по мере необходимости.

Наконец, документация нужна для того, чтобы понимать, почему появилась эта модель, а также ее основные методы. Она пригодится на случай, если работа над продуктом уже не ведется, а информация о том, почему он был создан, утеряна.

11.2.6. Тестирование

Прежде чем запускать модель машинного обучения в производство и давать клиентам возможность пользоваться ею, важно убедиться в ее работоспособности. При обучении модели МО полезно проверить ее выходные данные и обеспечение точности, но этого недостаточно, чтобы понять, будет ли модель работать. Ее необходимо протестировать, чтобы убедиться в ее способности обрабатывать без сбоев любые входные данные. Если вводимая в производство модель оттока имеет API, который принимает числовой идентификатор клиента в качестве входных данных, то что произойдет, если этот идентификатор будет пустым? Или если будет отрицательное число? Или слово «клиент»? Если в этих случаях API возвращает неожиданный ответ, это может быть не очень хорошо. А если неверные входные данные приводят к сбою API, это и вовсе может иметь катастрофические последствия. Таким образом, чем больше проблем удастся выявить заранее, тем лучше.

Тестировать можно разными способами, но при создании продукционной модели МО особенно важно юнит-тестирование — процесс тестирования каждого небольшого компонента кода, которое проводится с целью удостовериться, что система будет работать на практике. В случае с API для МО это часто означает проверку того, что каждая конечная точка API ведет себя должным образом в разных условиях. Это тестирование предполагает возможность получать в качестве входных данных очень большие и отрицательные числа или строки со странными словами. Каждый сценарий превращается в тест. Для модели машинного обучения, которая классифицирует анализ тональности текста, тест может быть таким: «При вводе “я люблю тебя” мы ожидаем, что ответ API будет положительным». Другой вариант теста может быть таким: если введено не целое число, например 27,5, то код вместо сбоя возвращает результат «невозможно вычислить».

Помимо конечных точек API можно тестировать отдельные функции в коде. Цель состоит в том, чтобы обеспечить стопроцентный охват, то есть каждая строка кода в API будет проверена на правильность работы. Каждый раз при развертывании модели проводится тестирование, и если оно не пройдено, то проблемы должны быть устранены.

Всегда есть соблазн отказаться от тестирования из-за нехватки времени, но обычно это единственный способ выявить серьезные проблемы до того, как модель будет представлена клиентам. Куча проверок может казаться бесполезной работой по сравнению с построением модели МО, но это чрезвычайно важно, и игнорировать этот этап никак нельзя.

11.2.7. Развертывание API

Если модель МО написана так, что ее можно запускать на ноутбуке, то преобразовать ее в API, который тоже будет запускаться на ноутбуке, не составит большого труда. К сожалению, иногда ноутбук приходится выключать или использовать для просмотра Netflix, поэтому нельзя оставить API работать на нем постоянно. Чтобы обеспечить постоянную и стабильную работу API, он должен располагаться где-то на сервере. Процесс перемещения кода на сервер для его выполнения — это то, что мы подразумеваем под развертыванием. Настройка сервера для постоянного выполнения кода предполагает немного больше работы, чем просто создание API.

Термин «сервер»

Слово «сервер» может звучать пугающе, словно это какой-то особый компьютер, неведомый простым смертным. На практике сервер — это обычный компьютер вроде того же ноутбука, но он работает где-то далеко от вас и без экрана. Вместо того чтобы идти к серверу и запускать систему, люди удаленно подключаются к другим компьютерам, имитируя нахождение рядом с ним. На серверах работают почти те же операционные системы, что и на ноутбуках, — Windows или Linux с небольшими изменениями. Если вы подключаетесь к серверу удаленно, интерфейс должен быть хорошо вам знаком: у него есть то же меню «Пуск» для Windows или терминал для Linux.

Когда вам говорят об использовании облачных сервисов, таких как Amazon Web Services (AWS), Microsoft Azure и Google Cloud Platform (GCP), то имеют в виду, что люди используют серверы, арендованные у Amazon, Microsoft и Google. Но то, что вы платите большой компании за эти услуги, не означает, что их компьютеры какие-то другие. Можете воспринимать их как дорогие ноутбуки.

Развернуть API на сервере можно двумя основными способами: запустить его на виртуальной машине (ВМ) или поместить в контейнер.

РАЗВЕРТЫВАНИЕ НА ВИРТУАЛЬНОЙ МАШИНЕ

Корпоративные серверы — это, как правило, чрезвычайно мощные и дорогие машины. Запускать их для одной задачи бессмысленно, потому что в большинстве случаев это излишне. Вместо этого сервер будет выполнять множество задач одновременно, но если всего одна задача приведет к сбою компьютера, а вместе с ним прекратится выполнение остальных задач, это было бы катастрофой. Виртуальные машины — решение этой проблемы, потому что они являются эмуляцией компьютеров. Большой и дорогой компьютер будет одновременно запускать множество эмуляций других компьютеров. Не страшно, если одна эмуляция выйдет из строя; другие продолжают работать. С ВМ почти во всех случаях можно обращаться так же, как с обычным компьютером; если вы войдете в одну из них, то не сможете определить, что это именно виртуальная машина, если, конечно, специально не зададитесь этим вопросом. Каждый раз при использовании AWS, Azure или GCP для доступа к компьютеру вы подключаетесь к виртуальной машине. Если вы попросите IT-отдел вашей компании приобрести для вас сервер, вам, скорее всего, также предоставят локальную ВМ.

Виртуальные машины хороши тем, что представляют собой эмуляцию: их можно легко включить или выключить. А еще можно делать снапшоты, чтобы вернуться к более ранней версии или запустить несколько копий одновременно. Снапшотом можно поделиться с кем-нибудь еще, кто будет работать с ВМ. Или можно вообще закрыть глаза и представить, что виртуальная машина — это обычный старый ноутбук (при условии, что вы пользуетесь ноутбуком с закрытыми глазами).

Поскольку ВМ — это обычный компьютер, то запустить код на ней совсем несложно: достаточно установить R или Python, необходимые библиотеки, скопировать на них этот код и, наконец, запустить его. Ровно то же самое вы будете делать для запуска API на своем ноутбуке! Если вам нужно внести изменения в API, просто скопируйте новую версию кода на виртуальную машину, а затем запустите его. Вы действительно можете запустить систему в производство всего за три шага:

1. Запустите виртуальную машину.
2. Установите программы и код, необходимые для запуска API модели МО.
3. Запустите свой API.

Учитывая количество людей, которые говорят о сложности создания продукционных систем, просто поразительно, насколько все можно упростить.

Одна из основных сложностей этого простого метода копирования и вставки кода на виртуальную машину и нажатия кнопки «Выполнить» заключается в том, что вам придется вручную перемещать код каждый раз, когда вы вносите изменения. Это достаточно трудоемкий процесс, где легко ошибиться. Можно запросто забыть переместить код или не вспомнить, какая его версия находится на виртуальной машине.

Непрерывная интеграция (continuous integration, CI) — это практика автоматической перекомпиляции кода каждый раз, когда он фиксируется в репозитории. Инструменты CI могут отслеживать репозитории git, отмечать, когда вносятся изменения, а затем перестраивать ПО на основе этой информации. Если вы используете R или Python, перекомпиляция, скорее всего, не понадобится, но процесс построения может включать такие шаги, как повторный запуск модульных тестов. *Непрерывное развертывание* (continuous deployment, CD) — это практика использования выходных данных инструментов непрерывной интеграции и их автоматического развертывания в продуктивных системах. CI/CD подразумевает совместное использование обоих методов.

Таким образом, инструмент CI/CD проверит ваш репозиторий на наличие изменений и в случае их обнаружения запустит процесс сборки (например, модульное тестирование), а затем переместит полученный код на виртуальную машину. Как дата-сайентисту вам не нужно беспокоиться о внесении изменений в ВМ: инструмент CI/CD сделает эту работу за вас. Самостоятельно настроить CI/CD — непростая задача, но если в компании есть команда разработчиков ПО, вполне вероятно, что они уже это сделали и вы можете пользоваться этим инструментом.

Еще одно преимущество виртуальных машин — возможность одновременного запуска нескольких ВМ. Если предполагается, что API будет получать большой трафик, то можно сделать копии виртуальной машины, запустить их одновременно и присвоить трафик машине случайным образом. Кроме того, можно отслеживать активность каждой ВМ, а также запускать и останавливать дополнительные копии по мере необходимости. Этот метод называется *автомасштабированием*. Он подходит для больших систем, но его довольно сложно настроить, и если вы оказались в ситуации, когда нужно использовать автомасштабирование, то вам наверняка потребуется помощь разработчика ПО.

РАЗВЕРТЫВАНИЕ В КОНТЕЙНЕР DOCKER

Настраивать и запускать виртуальную машину бывает довольно сложно. Поскольку каждая из них представляет собой эмуляцию компьютера, ее настройка не менее утомительна, чем настройка ПК. Вы должны установить каждую программу, изменить каждый драйвер и правильно все сконфигурировать. Кроме того, очень сложно задокументировать все необходимые шаги, а если кто-то повторяет этот процесс, то может легко допустить ошибку. Кроме того, виртуальные машины занимают много места, поскольку являются эмуляцией и должны содержать все, что делает обычный компьютер.

Docker решает эти проблемы. Согласно метафоре Майка Колемана (Mike Coleman), автора блога Docker (https://www.docker.com/blog/containers-are-not-vms), если считать, что сервер с виртуальными машинами — это район с домами, то контейнеры Docker — это набор квартир в отдельном доме. Несмотря на то

что каждая квартира полностью пригодна для проживания, все они делят между собой общие коммуникации вроде бойлера. По сравнению с ВМ контейнеры Docker настраиваются и запускаются проще и эффективнее.

Docker позволяет легко указать, как настраивается ВМ, и, имея общую спецификацию на разных машинах, вы можете использовать ресурсы совместно. Это существенно упрощает создание и поддержку продукционных систем, чем в случае с виртуальными машинами, поэтому Docker всецело захватил мир разработки ПО.

Чтобы разобраться с Docker, важно освоить три концепции:

- *Dockerfile* — текстовый файл, содержащий все шаги, необходимые для настройки моделируемой машины. Они могут включать вещи вроде «установить Python 3» или «скопировать сохраненный файл модели на машину». Большинство шагов в точности такие же, как и для команд Linux bash, поэтому, если они для вас привычны, dockerfile должен показаться вам знакомым.
- *Образ* Docker — результат, полученный после следования Docker шагам dockerfile для создания и хранения снапшота состояния компьютера.
- *Контейнер* создается, когда Docker берет образ и запускает его. Работающий контейнер можно подключить и использовать как обычный физический компьютер с программами и данными, указанными в образе.

Docker имеет множество преимуществ по сравнению с традиционными методами развертывания, но использовать его для развертывания модели МО в производство лучше всего, только если другие сотрудники компании уже работают с его контейнерами. Если это так, значит, кто-то из ваших коллег знает, как создать контейнер Docker, развернуть его и отслеживать его работу. В противном случае идею нестандартного развертывания моделей МО, скорее всего, не поддержат.

Если вы никогда раньше не развертывали код, то, возможно, вам будет проще начать с виртуальных машин, поскольку работа с контейнерами Docker устроена сложнее. Даже если вы не можете использовать Docker для развертывания моделей в производство, воспользуйтесь им для воспроизводимого анализа. Определенно стоит хотя бы немного поработать с этим инструментом, даже если сразу не очень понятно, пригодится ли он вам в работе.

11.2.8. Нагрузочное тестирование

Если модель будет применяться в нескольких системах или же одна система будет использовать модель множество раз одновременно, следует убедиться, что API выдержит нагрузку и не откажет. Этот сбой может произойти из-за того, что системе, в которой работает API, не хватает памяти, а также потому, что ей требуется слишком много времени для обработки каждого запроса и список очередности растет, или же по любой другой причине.

Самый простой способ убедиться, что этого не произойдет, — запустить *нагрузочное тестирование*, в котором вы одновременно делаете большое количество запросов к API и смотрите на реакцию. Обычно выполняется как минимум в два раза большее количество запросов, чем можно было бы ожидать. Если API обрабатывает эти запросы корректно, значит, все идет по плану. Однако если он выходит из строя, вы увидите, что нужно улучшить код, масштабировать систему или внести другие изменения.

11.3. Поддержание работоспособности системы

Даже после того, как вы успешно развернули и используете API, это еще не конец. (Всегда нужно что-то еще.) Либо вы, либо какой-нибудь другой сотрудник будет следить за работой API. В некоторых компаниях есть группа сопровождения (DevOps), которая занимается такими вещами. Даже если API работает нормально, вы все равно можете вносить изменения по другим причинам. В следующих разделах мы расскажем о трех важных моментах его обслуживания.

11.3.1. Мониторинг системы

Непрерывный мониторинг работы модели — отличная идея. Сколько запросов она получает каждый час? Верны ли прогнозы? Есть ли ошибки? Самый простой способ отслеживать эти показатели — включить в API логирование и телеметрию. *Логирование* — это запись данных о внутренних проблемах в программном средстве, например каждый раз, когда в модели возникает ошибка. *Телеметрия* — это регистрация возникающих событий, например каждый раз, когда делается запрос или конкретный прогноз. Кроме того, можно настроить оповещение при возникновении проблем.

Логирование можно организовать по простому принципу: скажем, сделать так, чтобы API записывал информацию в файл каждый раз, когда происходит событие. Затем для проверки логов можно просто войти в контейнер Docker или виртуальную машину. Телеметрия обычно включает отправку информации о событии на удаленный хост (например, централизованный сервер), чтобы информация со многих систем находилась в одном месте. Затем можно создать информационную панель, чтобы телеметрию можно было просматривать и контролировать в режиме реального времени.

Инструменты оповещения используются на случай, если что-то идет не так, чтобы сотрудники компании об этом узнали. Это могут быть автоматические электронные письма или сообщения в Slack, которые отправляются при наступлении определенного набора событий. Если в API модели настроено событие телеметрии о получении запроса, но в течение всего дня ни одного запроса не

поступило, можно отправить электронное письмо с оповещением о том, что система не получает трафик, хотя должна.

Эти разные системы мониторинга часто используются вместе, и компании пытаются стандартизировать их, чтобы все корпоративные API можно было отслеживать одинаково. Чем больше вы будете работать со стандартами вашей организации, тем полезнее будет ваше ПО.

11.3.2. Переобучение модели

Часто случается так, что в какой-то момент после запуска в производство модель начинает сбоить. Модели МО обучаются на данных, и со временем эти данные становятся менее актуальными. Например, модель машинного обучения для прогнозирования оттока клиентов может дать сбой, если у компании появятся покупатели из новых регионов. Если модель недостаточно эффективна, ее необходимо переобучить.

Самый простой вариант переобучения — повторить те же шаги, что и для первичного обучения модели, но загрузить при этом обновленные данные. Этот процесс предполагает загрузку данных в R или Python на ваш компьютер, повторный запуск сценариев и последующий запуск модели в производство аналогичным образом. Метод хорош тем, что если вы что-то уже делали, то сможете повторить. Многие крупные корпорации работают с продукционными системами МО именно так.

Вы можете усовершенствовать процесс, создав стандартный график, согласно которому будете выполнять работу. Вместо того чтобы пытаться уследить за метриками и интуитивно решать, когда модель нужно переобучать, возьмите за правило делать это каждые *n* недель или месяцев. Это избавит вас от необходимости гадать, когда же нужно совершить важное действие.

Что еще более важно, стандартный график позволяет автоматизировать процесс. Если у вас есть сценарий Python или R, который загружает данные, строит модель и затем где-то сохраняет ее, можно настроить систему на автоматическое выполнение этих действий по графику. Фактически можно сделать так, чтобы система переобучения самостоятельно запускалась в производство, так что вам не придется тратить время на такие вещи. Такая система может протестировать, работает ли заново обученная модель так же хорошо или лучше, чем предыдущая; если нет, то дата-сайентистам должно быть отправлено предупреждение. Подобные продвинутые процессы переобучения набирают обороты и поддерживаются облачными сервисами вроде AWS SageMaker.

Автоматические конвейеры переобучения стали очень популярным и не самым простым методом, но, в конце концов, пока вы в принципе переобучаете модель, все в порядке. Специалисты по данным наживают себе проблем, когда создают модель, развертывают ее в производство и постепенно перестают заниматься

ею, а модель тем временем работает все хуже и хуже. Не отслеживая производительность и не модернизируя модель, вместо помощи вы только мешаете своей работой. Следите за этим.

11.3.3. Внесение изменений

Если модель, запущенная в производство, хорошо проявила себя в бизнесе, вы неизбежно захотите внести в нее изменения, чтобы сделать ее еще лучше. Например, использовать больше наборов данных или изменить метод машинного обучения, чтобы повысить производительность API. Ваши коллеги также попросят добавить необходимые для них функции или сообщат о проблемах, которые были обнаружены при использовании модели.

Как мы уже говорили в главе 10, в контексте анализа такого рода изменения, пусть даже и одного элемента, могут приводить к реальным проблемам. Вопрос о том, стоит ли заниматься этой работой, даже если она интересна или кому-то кажется важной, весьма спорный. Если для повышения точности модели с 84 до 86 % потребуются три месяца, то вы просто потеряете это время, которое могли бы потратить на что-то другое. Или параметр, который кажется кому-то важным, может никак не повлиять на многих клиентов. Успешно развернутая в производство модель МО привлечет внимание многих людей, и ваша задача как датасайентиста, который помогал ее создать, заключается в том, чтобы не потратить время на улучшение модели напрасно.

11.4. В завершение

В этой главе рассматривается множество концепций развертывания моделей, с некоторыми из которых вы, возможно, были знакомы. Хотя не все темы могут иметь прямое отношение к вашей работе, не помешает иметь базовое представление о них на тот случай, если что-либо из описанного понадобится в будущем. В книгах и в интернете есть множество хороших ресурсов, где можно найти дополнительную информацию, особенно потому, что все эти темы тесно взаимосвязаны с программной инженерией. Поскольку Data Science продолжает меняться, они не потеряют свою актуальность, так что следует получше их изучить.

11.5. Интервью с Хизер Нолис, инженером МО в T-Mobile

Хизер Нолис (Heather Nolis) — инженер по машинному обучению в команде AI в T-Mobile, где она помогает запускать в производство модели R и Python миллионы раз в неделю. У нее есть степень магистра компьютерных наук и степень бакалавра нейробиологии и французского языка.

Что в вашей команде означает «инженер по машинному обучению»?
Я беру модели, которые создают дата-сайентисты, и преобразую их в продукты, которые поддерживает команда. В течение долгого времени специалисты по данным T-Mobile не особо напрягались: просто делали красивые модели и классный анализ, которые затем направлялись в отдел проектирования программного обеспечения для внедрения в производство. Идея заключалась в том, что так работа сможет иметь реальное влияние на бизнес, но инженерам было действительно сложно использовать ее, потому что между специалистами существовал огромный языковой барьер. Моя задача — стать связующим звеном, понять все, что вошло в анализ или важно для конкретной модели, а затем донести эту информацию до инженеров.

Каково было внедрять ваш первый фрагмент кода?
Свое первое развертывание я сделала в первую неделю в качестве разработчика ПО. В тот раз я работала с уже существующим продуктом, и сначала мне не было понятно, почему это сопряжено с риском. Когда пишешь код на компьютере, нормально запускать его 50 раз, чтобы проверить, как он работает. Но если при развертывании его в производственной среде хоть в одном его фрагменте найдется баг, это может повлечь за собой огромные последствия для компании, поэтому при выполнении кода, который не работает, неудобства возникают не только на вашем ноутбуке. В моем первом релизе, когда я была уверена, что все готово, мне фактически пришлось потратить еще часа три на интеграционное тестирование, прежде чем мне разрешили его выпустить.

Что произойдет, если в процессе производства что-то пойдет не так?
В самом начале я создала инструмент на основе Twitter, который рекомендовал бы вам ближайший магазин T-Mobile в социальных сетях. Я написала его на Node.js, который наша команда вообще не поддерживала, но решила: «Я сделаю все сама и покажу, что это возможно, и тогда кто-нибудь более квалифицированный сможет это разработать». Именно тогда я узнала, что «кто-то более квалифицированный сможет это разработать» никогда не происходит; мой код так и запустили в производство.

Мы выпустили продукт, при том что я была абсолютным новичком в Node, а получившийся код был, откровенно говоря, так себе. Он работал и был безопасен, но так как у меня был очень ограниченный опыт в производственной среде, я сомневалась, что все пройдет гладко. Другие инженеры нервничали, потому что это был язык, который мы еще не поддерживали. В течение следующих двух месяцев меня вызывали каждый раз, когда в какой-либо службе возникала небольшая техническая неисправность. И мне пришлось все время взаимодействовать с коллегами, потому что я рискнула выпустить что-то на другом языке на новой платформе; люди думали, что именно я была причиной всех странностей, которые происходили с программой.

Каждый раз, когда я получала уведомления все эти два месяца, я думала, что это моя проблема. Но на самом деле мой код никогда не сбоил в производстве! Я думаю, об этом нужно помнить: когда вы запускаете что-то в первый раз, есть шанс, что вы все испортите; с другими такое тоже случается. Дело не только в вашем коде. Не бойтесь, если что-то пошло не так.

Я могу сказать то же самое о запуске в производство: конечно, хотелось бы всегда делать самые красивые модели со всеми крутыми прибамбасами, но это не всегда приводит к созданию нужного продукта. В конце концов, нам приходится жертвовать многими из этих вещей, чтобы получить надежный код, который будет работать должным образом. Моя работа как инженера по машинному обучению — чувствовать этот баланс и стимулировать создание продукта.

Какой совет вы можете дать дата-сайентистам, работающим с инженерами?

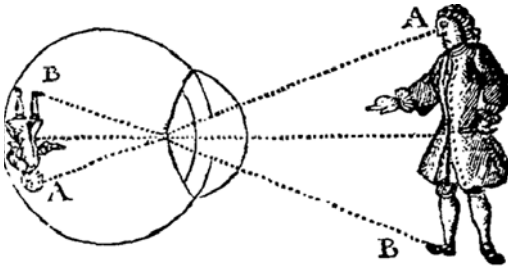
Два ключа к хорошей совместной работе — это понимать их язык и ценить то, что им важно. В первом случае помните, что то, что вы считаете нормальной фразой, для инженера МО звучит так, как если бы он пришел к вам домой и увидел огромный старый консольный телевизор с ЭЛТ. Они часто спрашивают себя: «Я вернулся в прошлое?» У меня есть любимый пример на эту тему, когда в T-Mobile в команде ИИ появился первый дата-сайентист. Как-то раз я спросила у него: «А нельзя ли просто запустить R-модель в производство как API?» Она переспросила: «Нужно, чтобы я запустила R как веб-сервер?» Я на секунду зависла, потому что слово «*веб-сервер*» для меня равносильно «привет, я из 80-х!» Меня это сбило с толку, хотя она имела в виду то же самое, что и я.

Специалистам по данным нравится строить точные модели. Мне как инженеру нравится запускать что-то в производство, чтобы другие могли с этим взаимодействовать. Единственное, что мне действительно дорого, — это работающий код. Если бы вы подошли к инженеру и сказали ему, что разработали API и создали документ, в котором указаны все входные и выходные данные, то тем самым доказали бы ему, что вам не безразличны их задачи.

Итоги

- Развертывание в производство — это практика создания моделей, работающих непрерывно.
- Помещение модели в REST API позволяет другим системам использовать ее.
- API могут быть развернуты на виртуальных машинах или в виде контейнеров Docker.
- Стоит внимательно изучить, как в компании работают с кодом, тестируют и развертывают продукционные системы.

12



Работа со стейкхолдерами

В этой главе

- Как работать с разными типами стейкхолдеров.
- Как взаимодействовать с людьми, не входящими в DS-команду.
- Как слушать внимательно, чтобы ваша работа использовалась наиболее эффективно.

Кажется, что работа дата-сайентистов будет в первую очередь связана с данными, но бóльшая ее часть вращается вокруг людей. Эти специалисты часами слушают, как их коллеги говорят о своих задачах и о том, как их можно решить с помощью данных. Они должны передать свою работу другим сотрудникам, чтобы те в дальнейшем могли использовать полученный в процессе анализа результат, или вверять им модели МО. В случае задержки проекта или недоступности данных им требуется выяснить у ответственных специалистов, каким должен быть следующий шаг.

В книге *Software Requirements* Карл Вигерс (Karl Weigers) и Джой Битти (Joy Betty) определяют термин «стейкхолдер» как «лицо, группу или организацию, которые принимают активное участие в проекте; на них влияет полученный результат, или они сами могут повлиять на него». Для дата-сайентиста стейкхолдерами могут быть бизнесмены из области маркетинга или разработки продукта, а также представители других сфер бизнеса, которые используют Data Science для принятия решений. Стейкхолдерами также могут быть инженеры, которые

применяют модели МО, созданные дата-сайентистами, для управления ПО или правильным сбором данных. В некоторых случаях стейкхолдерами являются руководители высшего звена. Заинтересованные стороны могут работать в разных департаментах компании, иметь разные потребности и вести себя по-разному.

В этой главе мы рассмотрим, чего следует ожидать от различных типов стейкхолдеров, с которыми вы столкнетесь в ходе проекта. Затем мы расскажем, как выстраивать с ними эффективные рабочие взаимоотношения и как вы должны думать, общаясь с людьми, не входящими в команду по работе с данными. Наконец, мы расскажем, как правильно расставлять приоритеты в задачах, которые поступают от стейкхолдеров.

12.1. Типы стейкхолдеров

У каждого стейкхолдера, который вам может встретиться во время проекта, есть свои опыт и мотивация. Хотя это может быть практически любой человек в зависимости от проекта, большинство из них попадают в одну из четырех категорий: бизнес, инжиниринг, высшее руководство и ваш руководитель (рис. 12.1).

Рис. 12.1 Типы стейкхолдеров, рассматриваемых в этом разделе

12.1.1. Бизнес-стейкхолдеры

Бизнес-стейкхолдеры — это люди из отделов маркетинга, обслуживания клиентов или продуктов, которые контролируют бизнес-решения. Именно они запрашивают анализ, который поможет принимать более обоснованные решения, или просят

предоставить модели МО, которые повысят эффективность. Люди, занимающие эти должности, имеют разное образование: у специалиста по маркетингу может быть степень MBA и опыт работы в рекламном агентстве, а руководитель отдела по работе с клиентами мог начать свой карьерный путь в качестве саппорта с дипломом местного колледжа, а затем продвинуться вверх и дорасти до руководителя. Разный опыт позволяет каждому по-своему взглянуть на работу с вами.

Обычно у бизнес-стейкхолдеров мало опыта работы в технической сфере. Они умеют работать с Microsoft Excel, и за некоторыми исключениями это и будет степень их аналитического опыта. Большинство бизнес-стейкхолдеров понятия не имеют, как использовать R или Python, а также не знают о достоинствах различных моделей МО. Но если они читали какие-либо статьи за последнее десять лет, то должны были много раз слышать о том, насколько ценными могут быть данные для принятия решений и насколько важен анализ данных в целом. Таким образом, бизнес-стейкхолдеры находятся в непростой ситуации: им приходится всецело полагаться на дата-сайентистов, которые предоставляют им крайне важную для принятия решений информацию или же предлагают инструменты машинного обучения, но при этом у стейкхолдеров нет достаточного количества технических знаний и они должны верить им на слово.

Часто бизнес-стейкхолдер активно вовлечен в DS-проект. Он помогает найти отправную точку и определить цель проекта. Он рассматривает промежуточные результаты и дает обратную связь от бизнеса, а также присутствует при завершении проекта, когда выдается окончательный анализ или развертывается модель. Поскольку именно бизнес-стейкхолдеры отвечают за ценность анализа, они должны быть постоянно вовлечены в процесс.

Ваша задача как дата-сайентиста — предоставить стейкхолдерам то, что им нужно для работы, например анализ, сводную таблицу или (в некоторых случаях) модель МО. Вы должны не только выполнить поставленную ими задачу, но и убедиться, что они понимают результат и доверяют ему. Если вы отдадите таблицу со сложной статистикой без объяснения, они ее не поймут и, следовательно, не смогут ею пользоваться. Став для стейкхолдеров надежным деловым партнером, вы даете им возможность использовать данные, а они открывают вам возможности для развития Data Science внутри организации.

Наиболее сложные ситуации с бизнес-стейкхолдерами обычно возникают, когда они отказываются принимать результаты анализа и отвечают что-то вроде: «Такого просто не может быть!» Если факты и предположения ставятся под сомнение, есть вероятность того, что дата-сайентистов отодвинут в сторону. В таких ситуациях лучшее, что вы можете сделать, — это объяснить стейкхолдерам, что именно и каким образом вы сделали. Обычно сомнения возникают из-за непонимания процесса; при обсуждении этапов, которые привели к полученному результату, возможно, вы решите изменить предположения в анализе.

12.1.2. Инженеры-стейкхолдеры

Команды инженеров отвечают за поддержку кода (и возможно, физического продукта), предоставляемого компанией, и если для этих продуктов требуются алгоритмы МО или анализ данных, они переходят в категорию стейкхолдеров. В некотором смысле работать с инженерами легче, чем с другими стейкхолдерами, потому что у них много общего с дата-сайентистами: у них также есть технический бэкграунд, полученный в университете или на онлайн-курсах.

Несмотря на обширные технические знания, у инженеров зачастую мало опыта в ключевых областях работы с данными. Хотя инженер ПО и пишет код, он обычно делает это под конкретную задачу, например ему нужно создать API, который отправляет запрос в определенную БД. Работа разработчика ПО не включает исследовательский компонент, который есть у дата-сайентистов, поэтому он не знает, что значит тратить недели на попытки понять данные и получить из них что-то полезное.

Как правило, инженеры взаимодействуют со специалистами по данным, когда в рамках инженерного проекта необходима модель машинного обучения. Такое взаимодействие чаще всего представлено в виде преобразования модели МО в API, развернутом в производство, который инженеры будут использовать в своей работе (см. главу 11). Инженеры полагаются на дата-сайентистов и специалистов МО, создающих надежный продукт с четкими входными и выходными данными, который не преподнесет неприятных сюрпризов при работе в производственной среде. Ваша задача — донести эти данные до инженеров-стейкхолдеров. Вы должны мыслить как инженер и попытаться понять, какой продукт лучше всего подходит для их нужд.

Инженерам также нужны дата-сайентисты, которые проводят анализ, помогающий улучшить их ПО. Специалисты по данным помогают определить приоритеты функций, диагностировать баги в инженерных системах и оценить производительность продуктов, ориентированных на клиентов, например это могут быть веб-сайты. В таких случаях инженеры находятся ближе к бизнес-стейкхолдерам, потому что им нужно получить информацию для принятия правильных решений.

Трудности с инженерами-стейкхолдерами обычно возникают из-за характера неопределенности Data Science. При разработке ПО у вас обычно есть цель настроить API или процесс, что вы и делаете. Перед вами стоят четкие задачи, связанные с проектным решением, и требования к тому, что продукт должен делать. А вот в Data Science ожиданий при создании продукта не так уж много. Не всегда понятно, какие данные в итоге понадобятся, потому что вы можете не знать заранее, что важно для модели. Неясно, каким будет результат, потому что он часто зависит от модели и ее производительности. И даже неизвестно, реализуема ли эта идея в принципе, потому что в какой-то момент может выясниться, что ни одна модель не является достаточно точной, чтобы соответствовать ожиданиям бизнеса.

Такая неопределенность на ранних этапах приводит к тому, что дата-сайентисты не могут многого обещать в самом начале, что крайне удивляет инженеров-стейкхолдеров. Таким образом, как специалисту по данным вам нужно особенно тщательно подбирать информацию о процессе, которую вы будете сообщать, чтобы инженеры не удивлялись внесенным изменениям. Заранее объясните и чаще напоминайте о том, что такое DS-процесс и как он выполняется. Если вы позволите инженерам разобраться в неоднозначности Data Science, они вряд ли будут удивляться.

12.1.3. Высшее руководство компании

Руководители компании имеют аналогичную квалификацию, что и бизнес-стейкхолдеры (или инженеры, если они руководят технологической организацией), но их сфера влияния гораздо шире. Директора, вице-президенты и главные должностные лица руководят корпорацией, и для этого им нужны данные. Дата-сайентистам часто поручают сортировку данных, чтобы получить информацию, необходимую руководителям для выполнения работы. Специалисты по данным также могут нести ответственность перед ними, если участвуют в крупномасштабном проекте, критическим компонентом которого является машинное обучение.

Руководители компании очень заняты, и у них нет времени на разбор деталей, которые их не касаются. Это приводит к огромным трудностям в коммуникации, и даже если у них и выдается свободная минутка, то она пролетает мгновенно. Когда вы встречаетесь с руководителем высшего уровня, он, как правило, хочет сразу перейти к делу и понять последствия. В этом есть смысл: это очень занятые люди, и чем меньше им приходится вникать в суть того, что кто-то пытается сказать, тем больше они могут сосредоточиться на принятии решений.

Руководители компании обычно взаимодействуют с дата-сайентистами, когда им нужны данные для принятия важного решения или если они хотят лучше понять какую-то часть компании. Иногда работу, которая выполняется для бизнес-стейкхолдеров или других заказчиков, показывают людям более высокого уровня в организации, прежде чем направить ее высшему руководству. В этих ситуациях анализ и отчет могут снова и снова совершенствоваться по мере продвижения вверх. В других случаях руководитель может запросить конкретный анализ или работу, и задача специалиста по данным — сделать так, чтобы для человека, который видит результат впервые, было все понятно.

В зависимости от размера и культуры организации дата-сайентистам, возможно, придется проделать большую работу, прежде чем поделиться результатами со стейкхолдером из высшего руководства. В некоторых компаниях есть специальные команды, которые проверяют результаты, чтобы убедиться, что они соответствуют бизнес-целям и убеждениям организации. В небольших фирмах специалисты

по данным могут проделывать работу непосредственно для руководителя. Независимо от компании работа всегда должна быть чистой и без ошибок.

Проблемы обычно возникают, когда работа сдается непонятной или неполной. Если руководитель не может понять, что ему принесли, у него не хватит терпения ждать, пока все прояснится само собой. Если он задает вопросы, на которые дата-сайентист не может ответить, то он может решить, будто на работу нельзя положиться. Если взаимодействие с руководителем не складывается и он не прощает ошибок, это может иметь серьезные последствия для команды.

С другой стороны, если руководителю понравятся результаты или он посчитает их ценными, то специалистам по данным это будет на руку. Став доверенным партнером руководителя, вы можете выиграть шанс использовать данные и машинное обучение в самых разнообразных целях и направлениях.

12.1.4. Ваш непосредственный руководитель

Ваш непосредственный руководитель иногда тоже выступает в роли стейкхолдера в зависимости от проекта, над которым ведется работа. Если он поручает вам задачу, постоянно проверяет ее и вносит предложения, то, по сути, является стейкхолдером в проекте. Менеджер хочет, чтобы проект был успешным, потому что (1) его работа — помочь вам добиться успеха, (2) он заинтересован в том, чтобы проекты под его руководством были успешными, и (3) проект может быть ориентирован на более широкие цели, заданные для команды.

В целом задача руководителя — направлять и наставлять сотрудника на протяжении проекта. Если у вас возникают трудности, вы должны иметь возможность обсудить их с менеджером, который поможет найти наилучшее решение. Руководитель может продвигать вашу работу настолько, насколько это возможно, рассказывая о ней другим специалистам, помогая интегрировать ее в существующие процессы и придумывая новые возможности ее усовершенствования.

Но руководитель также является стейкхолдером, потому что он полагается на вашу работу. Ему нужно, чтобы вы выполняли задачи наилучшим образом, потому что руководитель, в свою очередь, передает отчеты, модели и анализы, которые вы сдаете, другим специалистам. Таким образом, он выполняет двойную роль — человека, на помощь которого вы можете рассчитывать, и менеджера, для которого вы должны выполнять работу.

Поскольку прямой руководитель выполняет двойную функцию, все, что дальше написано в этой главе, также относится к нему. Основное отличие состоит в том, что со своим руководителем можно ослабить бдительность и быть немного более уязвимым. Например, ему можно сказать: «Ой, я изо всех сил пытаюсь закончить этот анализ, он дается мне очень непросто». Но вы едва ли можете сказать то же самое представителю высшего звена. Вероятнее всего, ваш менеджер будет отно-

ситься к вам по-людски и давать советы, тогда как другие стейкхолдеры являются просто заказчиками.

Относитесь к руководителям так же, как и к другим коллегам: давайте четкие апдейты, поддерживайте контакт и работайте на совесть. Если у вас возникнут трудности, сначала расскажите о них своему менеджеру, потому что он должен помочь вам и понять, когда для решения проблемы необходимо привлечь кого-то еще.

12.2. Работа со стейкхолдерами

Чтобы эффективно общаться со стейкхолдерами во время реализации проектов, необходимо придерживаться четырех основных принципов:

- понимать цели стейкхолдеров;
- постоянно общаться;
- быть системным;
- выстраивать взаимоотношения.

В следующих разделах мы подробно рассмотрим каждый из этих пунктов.

12.2.1. Понимание целей стейкхолдеров

У каждого сотрудника есть цели, которых он хочет достичь, приходя на работу каждый день. Эти цели определяются как должностью человека, так и его личными качествами, такими как амбиции и стремление к балансу между карьерой и жизнью. Например, ведущий инженер может сосредоточиться на завершении текущего проекта, чтобы получить повышение. Или же руководитель высшего звена может знать, что скоро уйдет из компании, и поэтому не хочет раскачивать лодку заранее. Эти цели определяют, чем сотрудники занимаются на работе и как они реагируют на действия коллег. Затянувшийся проект может раздосадовать инженера по продвижению, а руководитель, который ничего не хочет делать, только обрадуется.

Работая со стейкхолдерами, очень важно понимать их цели. Одинаковый анализ может считаться и хорошим, и плохим в зависимости от точки зрения. Рассмотрим анализ эффективности определенного продукта, продаваемого на веб-сайте компании. Предположим, что стейкхолдер — это человек, который управляет продуктом, а ваш анализ показал низкие продажи этого продукта в Южной Америке. Если стейкхолдер хочет представить продукт в лучшем свете, то такой результат анализа считался бы плохим, потому что он показал бы реальное положение дел. С другой стороны, если бы цель стейкхолдера заключалась в составлении каталога пользующихся спросом продуктов, то в таком случае информация о том, какой из продуктов лучше сократить, всегда полезна.

Работая со стейкхолдерами, нужно постараться как можно быстрее узнать их цели и мотивацию. Чем быстрее вы это сделаете, тем меньше вероятность сделать что-то не так. Понять мотивацию можно несколькими способами:

- *Спросите прямо*. Задавая прямой вопрос: «Что для вас важно?», вы фактически даете человеку возможность говорить открыто. Это не значит, что вы получите полную картину происходящего, но зачастую таким образом можно уловить самое главное. Кроме того, это совершенно нормальный вопрос во время вводного совещания.
- *Наведите справки*. Узнайте, работали ли ваши коллеги со стейкхолдером раньше. Задайте кому-нибудь из команды вопрос вроде: «Итак, расскажите мне об этом стейкхолдере. Чем он занимается?» Помните, что вы не должны сплетничать; не стоит принимать за чистую монету все, что вам сказал коллега, и необдуманно передавать это другим.
- *Судите о мотивации стейкхолдеров по их действиям*. Иногда поведение стейкхолдера может вполне четко указывать на мотивацию. Если вы предоставили анализ, который плохо характеризует один из продуктов, управляемых стейкхолдером, а сам он при этом занимает оборонительную позицию, это верный признак того, что продукт для него чрезвычайно важен. Обратная сторона этого метода заключается в том, что вам придется учиться в результате взаимодействия, поэтому можно легко ошибиться. Следовательно, постарайтесь делать выводы из предыдущих уроков.

Принимая участие в решении таких задач, вы сможете сформировать психологическую модель стейкхолдера. Как он отреагирует на тот или иной результат анализа или на несвоевременную сдачу модели? Если вы умеете заранее прогнозировать его поведение, то сможете эффективнее выстраивать общение.

Обратите внимание, что знание мотивации стейкхолдеров не означает, что вы должны ее удовлетворить. Хотя понимание целей помогает спрогнозировать реакцию, бывает и так, что ваши цели не совпадают с целями стейкхолдера, и тогда вам придется их игнорировать. Если вы стремитесь стать лучшим дата-сайентистом, то в ситуации, когда проделанный вами анализ показывает, что продукт работает плохо, в ваших же интересах предоставить подлинный результат и не скрывать реальные выводы. Зная потребности стейкхолдеров, можно облегчить себе работу.

Если вы понимаете, что должны сообщить новости, которым стейкхолдер будет явно не рад, вызывайте подмогу. Может ли ваш руководитель или более старший член команды прийти на помощь? Если кто-то возьмет эту задачу на себя, то тем самым освободит вас от любых политических последствий. От джуна все же не ждут блестящих навыков и знаний политики компании.

Если вам предстоит вести сложный разговор самостоятельно, попытайтесь представить его как взаимовыгодное сотрудничество. Подумайте, как встать на

сторону стейкхолдера. Новости действительно могут быть неприятными, но вы можете попытаться доказать, что получили такой результат не специально, пытаетесь посмотреть на ситуацию с его стороны и ищете возможные пути решения текущей проблемы. В этом случае беседа носит скорее деловой, а не технический характер; основной смысл в том, чтобы стороны пришли к взаимопониманию.

Ключевые показатели эффективности (KPI)

Ключевые показатели эффективности (KPI), а также цели и ключевые результаты (OKR) — это метрики, на которых сосредоточена команда или организация, поскольку они определяют ценность бизнеса. Команда интернет-магазинов может выбрать в качестве показателя, который они хотят увеличить, количество заказов в месяц. KPI полезны для дата-сайентистов, поскольку они обеспечивают явную количественную оценку целей команды. Если вы узнаете общий KPI, то сможете подстроить весь свой анализ и остальную работу так, чтобы это хорошо влияло на KPI коллег. Если анализ или метод не имеет отношения к KPI, команда, вероятно, не будет в нем заинтересована.

Не у каждой команды есть постоянные KPI, а в некоторых случаях они меняются или плохо сформулированы, но в ситуации, когда показатели все же заданы, лучше не игнорировать их. Зачастую это самый простой способ быстро понять цели стейкхолдера.

12.2.2. Постоянное общение

Дата-сайентисты часто переживают, что общаются либо слишком много, либо слишком мало. «Если я отправлю электронное письмо стейкхолдеру в третий раз за день, не будет ли это слишком?» Подобная мысль может прийти в голову, когда вы снова нажмете «Отправить». Или вы можете задаться вопросом: «Я давно не разговаривал с нашим стейкхолдером. Интересно, о чем он думает?» Или худший случай из всех: вы вообще упустили из виду, что надо держать стейкхолдера в курсе дела, и оставили его в неведении относительно продвижения проекта.

Дата-сайентисты почти всегда недостаточно коммуникабельны, стейкхолдерам же постоянно необходима обратная связь: электронные письма, совещания и звонки — единственные способы, с помощью которых они могут понять, что происходит с проектом. Без достаточного взаимодействия стейкхолдеры могут чувствовать себя не в курсе ситуации и переживать, что они плохо понимают происходящее. В условиях недостаточной коммуникации стейкхолдер может быть ошеломлен тем, насколько результат отличается от его ожиданий.

Специалист по данным должен информировать стейкхолдера о следующих вещах:

- Насколько проект укладывается в ожидаемые сроки. Если в начале предполагалось, что на сбор и очистку данных потребуется месяц, а затем еще один на построение модели, сообщите стейкхолдеру, актуальны ли эти сроки. В иде-

але дата-сайентист должен сообщать об изменениях и задержках по мере их появления. Плохой сценарий — это когда стейкхолдер ожидает завершения проекта, а у специалиста по данным впереди еще недели или месяцы работы, о чем он заранее не сообщил. Когда стейкхолдер об этом узнает, он может прийти в бешенство и вполне имеет на это право.

- Как продвигается проект: например, какие выводы сделаны в результате или какие части оказались труднее, чем ожидалось. Вопрос отсутствия доступа к БД потенциально может быть решен с помощью стейкхолдера. Информирование о том, где анализ идет хорошо, может помочь стейкхолдеру улучшить содержание проекта. Если вы считаете, что работа продвигается очень медленно, об этом тоже нужно сообщить. (В главе 13 есть дополнительная информация о неудачных проектах.)
- Помимо информации о продвижении проекта дата-сайентист должен постоянно держать стейкхолдера в курсе того, какая появилась информация для бизнеса и каковы дальнейшие планы. Специалисты по данным должны оценить, как проделанная работа повлияет на развитие проекта. Например, если в конце проведенного анализа обнаружилось что-то совершенно новое, необходимо сформировать для бизнеса набор рекомендаций о том, что делать с этим результатом.

Обычно наилучший способ наладить сбалансированное общение — просто сделать его основой работы проекта по умолчанию. Нет ничего лучше, чем регулярные встречи дата-сайентиста и стейкхолдера. Собираясь еженедельно или раз в две недели, вы обеспечиваете необходимый уровень коммуникации. Такой режим работы стимулирует: поставив отметку в календаре, вы заставляете себя делать что-то, чем можно поделиться на встрече. Вы должны каждый раз приходить подготовленным: с замечаниями об изменениях графика, заметками о том, что идет хорошо, а что плохо, с информацией о промежуточных результатах, которыми вы хотите поделиться, а также с предложениями относительно последующих действий.

Как дата-сайентист вы также должны выработать привычку при необходимости напрямую писать имейлы стейкхолдерам. Порой джуниоры боятся задавать вопросы руководству организации. Но в большинстве случаев стейкхолдеры будут рады ответить на вопросы, если это поможет лучше выполнить работу. Такова их роль в организации. Если вы переживаете, что человек занимает высокую должность и письмо нужно доработать, а ваши вопросы вызовут сомнения в вашей компетентности, сначала напишите своему руководителю: это его компетенция. В зависимости от стейкхолдера и проекта вы должны отправлять имейлы примерно раз в неделю.

Если в вашем проекте что-то внезапно меняется (возможно, ожидаемого датасета не существует) и вам нужно участие стейкхолдера, порой лучше по-

звонить или организовать внеплановую встречу. Этот вариант может подойти для быстрого решения вопросов, когда это необходимо. Единственный вопрос, который следует задать себе, прежде чем идти на этот шаг: «Действительно ли мне нужно мнение стейкхолдера по этому поводу?» Если что-то изменилось, но вы знаете, что делать дальше, не стоит впустую тратить чужое время. Но если участие необходимо, назначайте встречу. Частая ошибка начинающих дата-сайентистов — по умолчанию они считают, что встречи назначают им, а не они. Чем больше вы будете заботиться о продвижении проекта, тем лучше. Кроме того, этот опыт пригодится вам для будущей карьеры.

Метод и причины общения меняются в зависимости от типа стейкхолдера. В целом бизнес-стейкхолдеры обычно предпочитают встречи, которые обеспечивают им взаимодействие и управление проектом. Скорее всего, это не те люди, которые смогут помочь получить актуальные данные или решить проблему с техническими препятствиями. Инженеры обычно могут помочь с техническими вопросами, но в принятии решений о проекте или направлении работы они разбираются не больше вашего. Директора очень заняты и зачастую участвуют в проекте только вначале, при постановке целей, и в конце, при разборе результатов.

12.2.3. Будьте системным

Представьте себе ресторан в конце улицы. Однажды вы заказываете там фахиту, и вам тут же приносят вкуснейшую из тех, которые вы когда-либо пробовали. Через месяц вы снова заказываете фахиту, но на этот раз шеф-повар совершенно забыл добавить специи. Вы делаете заказ в третий раз: все было вкусно, но подача заняла больше часа — намного дольше, чем вы ожидали. Вы захотите снова прийти в это ресторан?

Компании процветают благодаря предоставлению надежного продукта, а вы как дата-сайентист являетесь мини-бизнесом в пределах своей организации. Стейкхолдеры — это ваши клиенты, и если вы плохо их обслуживаете, они перестанут просить вас о помощи. Один из способов обеспечить системность — стандартизировать работу.

В случае с анализом и отчетами вы можете существенно упростить процесс взаимодействия, создав для них единую структуру, чтобы делиться результатами. Если вы введете это за правило, то стейкхолдеры смогут сосредоточиться только на выводах работы. Вот несколько вещей, которые следует учитывать при стандартизации:

- *Как структурирован анализ*. Постарайтесь по возможности обеспечить определенный формат анализа. Начните с таких блоков, как «Цели» и «Данные» и завершите блоками «Выводы» и «Дальнейшие шаги». Так вы приучаете стейкхолдеров читать и анализировать ваши материалы.

- *Как проводится анализ*. Хотя следовать шаблону не всегда обязательно, практика показывает, что все идет более гладко, если для анализа вы используете один тип файла. Это может быть PowerPoint, PDF, HTML или что-то еще. Все файлы должны храниться в одном и том же месте. Можно настроить Dropbox, сетевую папку или другой инструмент для совместного использования. Убедитесь, что стейкхолдеры умеют им пользоваться; скорее всего, репозиторий GitHub не подойдет, хотя вы могли бы сделать его для себя, чтобы держать все под контролем версий.
- *Как оформлен анализ*. Этот момент может показаться не таким уж важным, но единый визуальный стиль может иметь большое значение. По возможности используйте одинаковые цвета и шаблоны (вам в плюс, если будете придерживаться корпоративного цвета компании).

Если вы создаете сводные таблицы, здесь работают те же правила, что и для анализа. Нужно сохранять единство стиля и формата для всех сводных таблиц и хранить их в общем месте, чтобы люди запомнили, где их можно найти.

И для API, и для программ МО системность заключается в дизайне продукта. По мере того как в DS-команде растет количество завершенных API и моделей, становится чрезвычайно сложно отслеживать, как работает каждый из этих продуктов. Чем более системный у вас подход, тем проще использовать API. Правила системности включают:

- *Системность входных данных*. То, как ваши модели и API принимают данные, должно максимально соответствовать одной и той же схеме. Например, все они принимают объекты JSON с одинаковыми именами параметров.
- *Системность выходных данных*. Структурирование выходных данных должно осуществляться с учетом того, как структурированы входные данные, а также того, как работают остальные API, созданные командой. Если модель принимает JSON в качестве входных данных, она должна возвращать JSON в качестве выходных.
- *Системность аутентификации*. Скорее всего, моделям и API понадобится определенная форма аутентификации в целях безопасности. Какой бы метод вы ни выбрали, он должен быть единым для максимально возможного количества API, особенно потому, что вы можете не уследить, какие учетные данные предназначены для конкретного API.

Такой порядок поможет не только стейкхолдерам, но и вам! Чем больше процессов вы сможете стандартизировать, тем меньше вам придется о них думать (и тем больше вы сможете сосредоточиться на интересных задачах). Чем больше вещей группа дата-сайентистов может стандартизировать, тем легче им будет передавать работу коллегам. Стандартизация хороша для всех.

Элизабет Хантер (Elizabeth Hunter), старший вице-президент по реализации технологической стратегии в T-Mobile: управление взаимоотношениями

Взаимоотношения играют важную роль в любом деловом взаимодействии, но иногда их упускают из виду. Люди подсознательно ищут социальные и эмоциональные сигналы, чтобы установить взаимосвязь, и когда они их находят, то становятся более открытыми для новых идей и опыта. Такая связь с человеком создает благоприятную атмосферу для передачи любой информации. Установить хорошие взаимоотношения — не такое уж линейное дело по сравнению с другими рабочими задачами. Некоторым удается быстро установить связи благодаря своей доброжелательности. Другим для этого требуется гораздо больше времени, многочисленных взаимодействий и личного общения на глубокие темы.

Со временем я обнаружила, что бóльшая часть моего успеха в карьере зависит от взаимоотношений, для выстраивания которых я нашла время, будь то чья-то поддержка на важной встрече, общение с руководителем, выслушавшим одну из моих идей, или кем-то, предложившим мне новые возможности. Огромная часть моего карьерного роста связана с упорным трудом и доказыванием ценности моей работы. Но, когда я установила личные связи, у людей появилось лучшее представление обо мне, а также о том, чего от меня можно ожидать, сколько свободы действий мне можно предоставить и насколько мне можно верить на слово, вместо того чтобы просить предъявить доказательства.

Мне как интроверту было сложнее, чем другим, выстраивать взаимоотношения; приходилось над этим работать. Раньше я проводила ряд небольших экспериментов: формировала гипотезы на основании своих знаний или наблюдений о человеке и о том, как с ним можно общаться, потом проверяла, работает ли моя гипотеза, затем на основании новой информации я корректировала ее и следовала ей при взаимодействии с человеком до тех пор, пока не выясняла, чем он живет и что собой представляет, и начинала хорошо к нему относиться. Это не означает, что вы должны изменить себя ради других, но эмпатия по отношению к тому, что нравится другим людям, приносит на удивление хорошие плоды.

12.3. Расстановка приоритетов

Дата-сайентисту часто приходится решать, над какой задачей работать в первую очередь. Несмотря на то что в некоторых командах есть руководитель проекта, который ставит задачи каждому специалисту, не стоит оставаться в стороне: вы также можете советовать, чем заниматься дальше. Работа может сильно различаться по тематике и масштабам и исходить от разных стейкхолдеров. Задачи можно разделить на три категории:

- *Разовые задачи, которые поступают непосредственно от стейкхолдеров.* Обычно они представляют собой небольшие запросы, например: «Составьте график продаж за определенное время». Они часто бывают срочными, и, по-

скольку они не занимают много времени, трудно ответить на такой запрос отказом. Но каждая такая задача отвлекает от более важной работы, и по мере накопления запросов становится все труднее работать продуктивно.

- *Долгосрочные проекты для бизнеса*. Эти проекты составляют основу работы дата-сайентиста. К ним относятся создание сводных таблиц, подробный анализ и настройка моделей для развертывания в производство. Эти задачи, как правило, очень важны, но, учитывая, что на их выполнение могут уйти недели или даже месяцы, они не всегда являются срочными.
- *Идеи, которые, по вашему мнению, являются перспективными*. Учитывая природу Data Science, они, как правило, носят более технический характер. Например, это может быть создание модели МО для прогнозирования обращения клиента в службу поддержки. В эту категорию также входят задачи по оптимизации, например создание функций или даже библиотеки для более быстрого решения общих вопросов. Если на выполнение процесса вручную каждую неделю уходит несколько часов, можно автоматизировать задачу, не принося компании прямой выгоды, но косвенно обеспечивая возможность делать больший объем работы. Никто об этом не просит, но такая работа важна.

Определить, что самое важное, а что можно отодвинуть на второй план, не всегда просто, особенно если к вам обратились сразу несколько человек. В то же время работа, которая в приоритете у стейкхолдеров, может не иметь значения для бизнеса в целом. В редких случаях дата-сайентист может отклонить их запросы, ведь обычно именно они определяют направление развития бизнеса. Все это создает среду, в которой то, как вы расставите приоритеты, может существенно повлиять на бизнес; это также дает вам меньше свободы действий.

Многие специалисты по данным испытывают трудности в этом вопросе. Когда им поступают задачи от стейкхолдеров, естественно, они хотят угодить им, а не разочаровать. Кроме того, такие запросы могут быть интересными с профессиональной точки зрения. Однако попытки справиться со всем сразу нереальны, ведь поток задач никогда не иссякает. Кроме того, ответ на один вопрос часто приводит к появлению новых, поэтому так вы скорее создадите дополнительную работу, а не снизите ее объем.

Когда вы обдумываете возможные задачи, над которыми стоит поработать, задайте себе следующие вопросы:

- *Будет ли эта работа иметь значение*? Повлияют ли результаты этого анализа на компанию? Могут ли они изменить решения? Увеличит ли эта модель МО прибыль?
- *Привнесет ли эта работа что-то новое*? Вы бы стали применять существующий процесс снова и снова или попробовали бы что-то другое?

Ответы на эти вопросы разбивают работу на четыре вида:

- инновационная и полезная;
- не инновационная, но полезная;
- инновационная, но не полезная;
- не инновационная, не полезная.

В следующих нескольких разделах мы подробно рассмотрим каждую из этих комбинаций.

12.3.1. Инновационная и полезная работа

Инновационная работа, которая меняет бизнес, — именно то, чему хотят посвятить себя большинство дата-сайентистов. Например, это может быть проект, где используются инвентаризационные данные, к которым такой специалист ни разу не прикасался, и новейшая модель МО для оптимизации заказа продукта, что позволит сэкономить компании миллионы долларов. Это те проекты, которые в лучшем случае дают вам возможность засветиться в журналах вроде Harvard Business Review или Wired.

К сожалению, в эту категорию попадают не многие проекты компании. Чтобы они существовали, должен быть выполнен ряд условий:

- Для эффективности методов анализа должно быть достаточно данных.
- В данных должен быть интересный сигнал, который модели могут уловить.
- Проект должен затрагивать крупную или достаточно важную часть бизнеса, чтобы изменения могли повлиять на прибыль (поэтому проект не направлен на оптимизацию складских запасов маркеров на водной основе).
- Задача должна быть новой, достаточно сложной или уникальной. Набор задач в компании, которые соответствуют всем этим категориям, чрезвычайно мал.

Эти проекты хороши тем, что вызывают интерес как у стейкхолдеров, так и у дата-сайентистов. Первые чувствуют себя отлично, потому что ясно видят ценность проекта. Вторые стремятся опробовать новые методы на новых данных и увидеть результаты. Если вам встретится проект из этой категории, сделайте все возможное, чтобы его развивать. Такие проекты могут стать определяющими для карьеры, но из-за того, что к ним предъявляется так много требований, они встречаются крайне редко и не всегда успешны.

12.3.2. Не инновационная, но все же полезная работа

Эти проекты не являются инновационными, но они меняют бизнес. Например, это может быть весьма обыденный анализ данных, который убеждает команду запустить продукт. Часто это убеждение сводится к доказыванию того, что все и так знают;

это не особенно новаторский подход, но он работает. В инженерном отношении эти проекты могут включать модель, которая уже была развернута в одном подразделении компании, и теперь ее нужно перенести в другое. Еще один вид работы, попадающий в эту категорию, — оптимизация задач, на решение которых требуется много времени. Такая работа не инновационна, но она оптимизирует бизнес.

Хотя подобные задачи не такие крутые, важно то, что они идут на благо бизнесу. Помогая стейкхолдерам увидеть ценность работы в области Data Science, можно заручиться их поддержкой. Если у вас будут сторонники, то в следующий раз, если проект выйдет за рамки бюджета или не принесет золотые горы, люди с большей вероятностью сохранят в вас веру. По возможности старайтесь браться за такие проекты.

В конце концов, работа дата-сайентиста состоит в том, чтобы приносить пользу компании, а не заниматься только самыми интересными вещами. Ценным качеством является умение терпеть такие важные, но скучные задачи. Однако если работа целиком и полностью состоит из подобных неинтересных проектов, которые ничему не учат, вполне уместно начать поиск чего-то другого. Совершенно нормально при расстановке приоритетов учитывать личный интерес и удовлетворенность работой; просто убедитесь, что это не единственный параметр.

12.3.3. Инновационная, но не полезная работа

Такая работа инновационна, но бесполезна для бизнеса: например, исследование новых теоретических алгоритмов и методов анализа, которые вряд ли будут использоваться. Подобные проекты могут быть оторваны от реальности; люди могут месяцами или годами корпеть над задачей, не взаимодействуя при этом с другими командами, а в конечном итоге продукт не пригодится. Эти проекты могут отнять огромное количество времени и обойтись в миллионы долларов, не показав при этом полезный результат. Несмотря на эти факты, дата-сайентисты летят на такие проекты как мотыльки на свет.

Как правило, старт подобных проектов дается внутри DS-команды. Такие проекты сосредоточены на том, что интересно методологически, а не на пользе для бизнеса. Специалист по данным может прочитать научную статью, в которой излагается новая теоретическая методика, и убедить остальных, что им стоит опробовать ее на собственных данных. Через полгода становится очевидно, что метод не так хорош, как было описано, а даже если бы он оправдал ожидания, то особой потребности в результатах, которые мог бы обеспечить алгоритм, у бизнеса все равно не было. Хуже того, дата-сайентист прочитал новую статью и ситуация повторяется.

Часто стейкхолдеры даже не знают об этих проектах. В лучшем случае они замечают, что некоторые специалисты по данным очень заняты работой над чем-то,

что звучит очень сложно, но никто не говорит, что это такое. Дата-сайентисты порой считают, что по завершении проекта люди смогут найти ему применение. На практике, если вы не видите применение проекту сразу, стейкхолдеры, вероятно, тоже его не найдут. Постарайтесь не зацикливаться на бесполезной работе, которая поставит под сомнение вашу ценность.

12.3.4. Не инновационная, не полезная работа

К сожалению, огромная часть работы, которую ждут от дата-сайентистов, ни инновационна, ни эффективна. Самый классический пример — это регулярно обновляемый, неавтоматизированный отчет, на создание которого уходит много времени, но никто его даже не открывает. Такая работа требует кучу времени и усилий, но если она делается для нескольких стейкхолдеров, то ни один из них не захочет взять на себя ответственность и сказать, что так делать больше не нужно. Поскольку со временем количество отчетов только растет, в конечном итоге DS-команда может оказаться перегружена.

Хотя отчетность — это один из видов работы, который не может быть ни инновационным, ни эффективным, в эту категорию может попадать множество небольших разовых запросов. Руководители, которым нравятся данные и диаграммы, могут неоднократно обращаться с такими просьбами, как: «Сделайте мне график продаж в Европе по неделям» или «Найдите мне продукт, заказы на который упали больше всего за последние 12 недель». Ни один из этих запросов не может быть особенно сложным, но вместе они требуют больших ресурсов и не представляют особой ценности для бизнеса.

Это непростые ситуации, потому что здесь нет однозначных решений. Можно попробовать автоматизировать отчеты и процессы, которые отнимают много времени, но задача сама по себе требует много времени, и вы можете лишь незначительно ее оптимизировать в зависимости от используемой технологии. Если стейкхолдеры на высшем уровне повторно обращаются к вам с запросом, трудно сказать «нет», не поставив под угрозу репутацию своей команды.

Несмотря на эти сложности, вы как дата-сайентист обязаны настаивать на том, чтобы ваше время использовалось целесообразно. Поговорите с руководителем или стейкхолдером: если вам поступает много таких задач, вы должны дать понять, что они могут не стоить затраченного на них времени. Скорее всего, они уже знают, что эта работа не особенно полезна, а если вы предложите варианты оптимизации, то вам, скорее всего, пойдут навстречу, так как текущее положение дел перестанет их устраивать. Если нет, то иногда лучшее, что вы можете сделать, — сначала попытаться самостоятельно улучшить процессы, а затем продемонстрировать результат.

12.4. В завершение

Работа со стейкхолдерами — непрерывный процесс на протяжении всего проекта. Вы должны понимать их потребности и цели просьб. Проект запускается благодаря запросу стейкхолдеров, но то, что они требуют, скорее всего, изменится в ходе самого проекта, и вы обязаны не отставать от изменений. Чем больше проект будет соответствовать требованиям стейкхолдера, тем меньше вероятность неудачи. В главе 13 мы поговорим о том, что происходит, если DS-проект провалился, например в таких случаях, когда общение со стейкхолдерами не заладилось.

Сэм Бэрроуз (Sam Barrows), дата-сайентист в Airbnb:
Постройте диалог вокруг запросов

Одним из ценных инструментов работы со стейкхолдерами является умение вести диалог. Коллеги могут часто просить вас выполнить какие-нибудь задачи. Вместо того чтобы сразу соглашаться или отказываться, спросите, почему возник запрос. Какую бизнес-задачу он решит? Есть ли более подходящий способ достичь желаемого результата? Понимая мотивы поступающих запросов, вы с большей вероятностью будете выполнять значимую работу.

Эта стратегия предполагает переговоры, в которых учитываются общие интересы, а не только сиюминутные потребности. Сиюминутные потребности — это запросы, которые вы получаете, тогда как основные интересы — это бизнес-мотивация, стоящая за ними, а также цели команды по работе с данными.

12.5. Интервью с Сейд Сноуден-Акинтунде, дата-сайентистом в Etsy

Сейд Сноуден-Акинтунде (Sade Snowden-Akintunde) работает в Etsy, где специализируется на разработке и анализе экспериментальных программ для улучшения покупательского опыта международных потребителей. В область ее специализации входят A/B-тестирование и экспериментальные программы, внедрение надежных методов обработки и масштабирование инфраструктуры данных.

Почему важно уметь взаимодействовать со стейкхолдерами?

К сожалению, не важно, насколько вы умны, если не можете объяснить идею стейкхолдерам без технического образования. В конце концов, многими компаниями руководят люди, которые могут не иметь того же уровня технических знаний, как у вас. Вы должны уметь общаться с ними так, чтобы они доверяли вам и позволяли отстаивать свое мнение, если это необходимо. Умение взаимодействовать со стейкхолдерами является, пожалуй, одним из самых важных качеств, но часто ему уделяется наименьшее внимание.

Как вы научились взаимодействовать со стейкхолдерами?

Методом проб и ошибок: у меня были разные ситуации — удачные и не очень, и я делала выводы. Думаю, самое важное, что мне удалось усвоить, — это необходимость как можно раньше начать диалог и повторять свои слова, чтобы убедиться, что вас понимают. В начале своей карьеры я думала, что если я что-то однажды сказала и кто-то со мной согласился, значит, человек точно знал, о чем я говорю, но порой люди могут даже не осознавать, что не понимают сути.

У вас когда-нибудь возникали трудности со стейкхолдерами?

Будучи разработчиком экспериментов, я поначалу боялась отстаивать себя и свою точку зрения. Другие люди проводили эксперименты, которые я считала бессмысленными, но я молчала. Затем, когда эксперименты закончились, я попыталась проанализировать их, но мне было сложно интерпретировать результаты из-за особенностей планирования эксперимента. Мне стоило бы поговорить со стейкхолдерами с самого начала и сказать, как я могу улучшить анализ и что получится, если планирование будет правильным. Я поняла, что если хочу сделать работу качественно, то обо всех нюансах следует договариваться «на берегу».

Какие ошибки чаще всего допускают молодые специалисты по данным?

Я думаю, что джуниоры полагают, будто другие автоматически осознают ценность их работы. Это мнение особенно распространено среди дата-сайентистов с академическим образованием. Нас поглощает все, к чему требуется суперосновательный или научный подход. Несмотря на то что это важно в академических кругах, один лишь упорный труд не обязательно заставит других признать ценность вашей работы. А вот то, как вы преподносите информацию, позволит людям оценить результаты как что-то стоящее.

Всегда ли вы пытаетесь объяснить техническую часть Data Science?

Это зависит от того, насколько это интересно стейкхолдеру. Я работала с менеджерами проектов, которые не хотели вникать в технические вопросы. Если бы я просто сказала: «Сейчас это не работает», они бы поверили. Я также работала с руководителями проектов, которые хотели знать каждую деталь, и когда я посвящала их в подробности, они, как правило, были немного в шоке. Некоторые хотят, чтобы вы регулярно отчитывались о происходящем, и их совершенно не смущает тот факт, что они ничего не смыслят в этом вопросе: им просто нужно быть в курсе событий. Так что я просто даю им то, чего они хотят.

Какой совет вы дадите начинающим дата-сайентистам?

Я думаю, что люди выбирают техническую стезю, потому что считают, будто смогут сосредоточиться только на логическом элементе и исключить человеческий фактор. Но это совершенно не так. Рассматривая вариант карьеры в Data Science,

стоит действительно задуматься, готовы ли вы отодвинуть свое эго на второй план, чтобы выстраивать общение с другими людьми и хорошо выполнять свою работу. Нет ничего проще, чем сказать: «Я хочу научиться строить эту модель, изучить A/B-тестирование и понять все эти технические вещи». Это здорово, но именно гибкие навыки помогут вам продвинуться дальше.

Итоги

- Есть разные типы стейкхолдеров с разными потребностями.
- Выстраивайте взаимоотношения со стейкхолдерами, чтобы они могли положиться на вас.
- Поддерживайте постоянный контакт и держите стейкхолдеров в курсе насчет сроков завершения работы и трудностей с проектами.

Материалы к главам 9–12

Книги

Beautiful Evidence, Eduard Tufte (Graphics Press)

Эдвард Тафти — легенда в области визуализации данных; его книги наполнены подробными инструкциями о том, как продумывать графики и таблицы до мелочей. У него есть и другие книги: вы можете взять их в наборе или, что еще лучше, пройти один из его однодневных курсов, которые он проводит на выездных турах. Но учтите: его советы иногда носят академический характер. Практически невозможно следовать всем его рекомендациям так, чтобы оставалось время на что-либо еще, кроме визуализации.

Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures, Claus O. Wilke (O'Reilly Media)

Если Эдвард Тафти дает обзор академического подхода к визуализации, то Уилке предлагает прикладную версию. В его книге описано, как работать с визуализацией на повседневной основе. Когда стоит использовать коробчатые диаграммы? Так ли плохи круговые диаграммы, как о них говорят? Эта книга поможет вам в принятии подобных решений.

The Pyramid Principle: Logic in Writing and Thinking, Barbara Minto (Trans-Atlantic Publications)

Эта книга давно не переиздается (хотя вы можете найти подержанный экземпляр), но до сих пор считается основополагающей работой по улучшению коммуника-

ции. Минто объясняет, как структурировать отчет или презентацию так, чтобы они были понятны аудитории, а также дает важные рекомендации: например, как упорядочить контент, чтобы он был логически последовательным, а не просто повторял порядок, в котором вы его создали. Минто — бывший консультант престижной консалтинговой фирмы McKinsey, и в книге много уроков, которые консультанты осваивают в своей работе.

The Design of Web APIs, Arnaud Lauret (Manning)

Принято учиться разрабатывать API самостоятельно: со временем вы наработаете достаточно опыта, чтобы дизайн стал приемлемым. Эта книга как бы показывает вам кратчайший путь. Она начинается с описания того, что такое API и как они структурированы, затем описывает их разработку и лучшие практические варианты. Она даже охватывает такие темы, как документация OpenAPI, так что вы сможете писать общедоступные инструкции для своих API.

Amazon Web Services in Action, 2nd ed., Michael Wittig and Andreas Wittig (Manning)

Azure in Action, Chris Hay and Brian H. Prince (Manning)

Google Cloud Platform in Action, JJ Geewax (Manning)

В этих трех книгах рассказывается, как использовать Amazon Web Services, Microsoft Azure и Google Cloud Platform соответственно. Когда вы научитесь развертывать модели МО в производство, вам понадобится место для их размещения, и эти три облачных провайдера — основные из доступных. Вы можете выбрать ту платформу, которая кажется вам наиболее удобной, а затем прочесть соответствующую книгу, чтобы изучить основы.

Difficult Conversations: How to Discuss What Matters Most, Douglas Stone, et al. (Penguin Publishing)

Общение — это всегда непросто, а когда тема важная или люди глубоко увлечены, оно становится еще сложнее. Эта книга посвящена разговорам, которых люди обычно избегают, и перечисляет отличный набор навыков для дата-сайентистов, потому что им часто приходится демонстрировать коллегам неприятные результаты.

Getting to Yes: Negotiating Agreement Without Giving In, Roger Fisher, William L. Ury, and Bruce Patton (Penguin Publishing)

Работа дата-сайентиста предполагает частые переговоры: от убеждения команды предоставить доступ к данным до обоснования руководителю своих выводов. Умение успешно убеждать и вести переговоры в такие моменты может быть важнее для вашего успеха, чем любые технические навыки. Эта книга — отличный ресурс, чтобы научиться вести переговоры со стейкхолдерами и получать желаемые результаты.

Software Requirements, 3rd ed., Karl Wiegers and Joy Beatty (Microsoft Press)

Иногда бывает сложно определить, что необходимо для проекта, в том смысле, который подразумевает бизнес. В этой известной книге рассказывается, как создавать техническое задание и управлять им в ходе проекта. Хотя сбор требований — не самая привлекательная часть Data Science, он может упростить либо нарушить реализацию проекта.

Блоги

«R in Production», Jacqueline Nolis and Heather Nolis

http://mng.bz/YrAA

В этой серии из трех частей рассказывается о создании API-интерфейса в R с помощью программного пакета plumber, его развертывании в виде Docker-контейнера и доведении его до рабочего состояния. Предоставляемый Docker-контейнер R с открытым исходным кодом используется T-Mobile.

«Advice for new and junior data scientists: what I would have told myself a few years ago», Robert Chang

http://mng.bz/zlyX

В этом популярном посте Роберт Чанг, наш собеседник из главы 1, излагает шесть основных принципов, которые он усвоил на своем пути к должности старшего дата-сайентиста в Airbnb. Самостоятельный путь к осознанию этих важных идей может занять годы, так что воспользуйтесь готовым опытом и начните применять его сейчас.

«Data Science foundations: know your data. Really, really, know it», Randy Au

http://mng.bz/07Pl

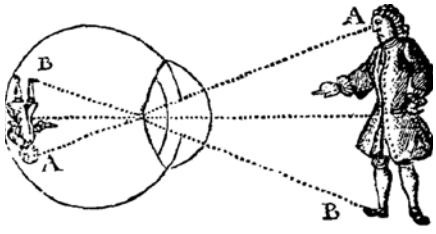
Рэнди Ау, наш собеседник из главы 2, дает этот совет каждому новичку в Data Science: «Разберитесь в своих данных: откуда они получены, из чего состоят, что они значат. Все начинается с этого». В этом посте он расскажет, как понять свои данные, начиная со структуры и заканчивая принятием решений по их сбору.

«How to work with stakeholders as a data scientist: what I would have told myself when I started», Sam Barrows

http://mng.bz/KEPZ

Мы поделились первым из семи советов Сэма по продуктивной работе со стейкхолдерами в примечании к главе 12 («Постройте диалог вокруг запросов»), но остальные шесть тоже стоит прочитать.

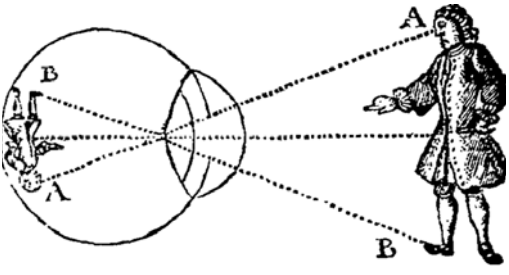
Часть 4 Как подняться по карьерной лестнице в Data Science



Заключительная часть этой книги — это материал, который нужно использовать после того, как вы освоитесь с должностью дата-сайентиста; здесь мы расскажем о том, что будет дальше. Темы этой части касаются каждого специалиста по данным, но о них нечасто говорят. Можно подумать, что если у вас есть стабильная работа в Data Science, то вы добились своего; но всегда есть к чему стремиться. Цель этой заключительной части — рассказать, как вырасти из джуниора в синьора и не только.

В главе 13 объясняется, что делать, если проект провалился. Эта тема крайне важна для опытных дата-сайентистов, потому что вы наверняка столкнетесь с неудачами на протяжении карьерного пути. Глава 14 посвящена вступлению в сообщество специалистов по данным: здесь вы узнаете, как писать статьи в блогах и участвовать в конференциях. Хотя это и не обязательно, но членство в сообществах может принести пользу в виде сети контактов и предложений о новой работе. В главе 15 мы поговорим о непростой задаче — как уволиться с наименьшими потерями. В последней, 16-й главе этой книги обсуждаются некоторые из основных вариантов карьерного роста после должности старшего дата-сайентиста, например возможности стать менеджером или техническим руководителем.

13



Если DS-проект провалился

В этой главе

- Почему DS-проект может потерпеть неудачу.
- Что делать, если это произошло.
- Как справиться с негативными эмоциями, возникшими из-за провала.

Большинство проектов по работе с данными и их анализу сопряжены с высоким риском. Вы пытаетесь спрогнозировать то, чего никто раньше не прогнозировал, оптимизировать то, что никто не оптимизировал, или понять данные, на которые раньше никто не смотрел. Независимо от того, чем именно вы занимаетесь, вы делаете это первым; работа почти всегда носит исследовательский характер. Поскольку дата-сайентисты постоянно занимаются чем-то новым, вы неизбежно попадете в ситуацию, когда обнаружите, что то, на что вы рассчитывали, просто невозможно. Смиритесь с мыслью, что идея оказалась безуспешной. Неудача — это душераздирающий и мучительный опыт; вы расстроены и хотите бросить Data Science и вообще уйти из этой области, хлопнув дверью.

В качестве примера рассмотрим компанию, создающую модель МО, которая будет рекомендовать продукты на веб-сайте. Вероятный ход событий начинается с нескольких совещаний, на которых команда дата-сайентистов убеждает руководителей в том, что этот проект — хорошая идея. Команда считает, что с помощью

информации о клиентах и их транзакциях можно спрогнозировать, какие еще продукты захотят приобрести покупатели. Руководители разделяют идею и дают проекту зеленый свет. У многих других организаций есть модели, которые кажутся простыми, поэтому все должно сработать.

К сожалению, как только команда начинает работу, наступает жестокая реальность. Может внезапно выясниться, что из-за недавней смены систем в компании данные о транзакциях доступны только за последние несколько месяцев. Или, может быть, команда проводит эксперимент и обнаруживает, что клиенты, которые видят движок рекомендаций, покупают не больше тех, кто его не видит. Подобные проблемы накапливаются; в конце концов, команда в смятении отказывается от проекта.

В этой главе мы определяем *провалившийся* проект как не достигший поставленной цели. В случае с анализом проект может считаться провалившимся, если он не отвечает на бизнес-вопрос стейкхолдера. В случае с задачей машинного обучения проект может провалиться, если модель не удается развернуть или она не будет в итоге работать. Проекты могут потерпеть неудачу по многим причинам.

Дата-сайентисты обычно не любят говорить о провале проектов, хотя это случается очень часто. Если проект не удался, специалист по данным может воспринимать это крайне болезненно. Это произошло с вами, вы можете подумать: «Будь я профессионалом, этого бы не случилось». Никто не любит делиться сомнениями в собственных способностях.

По своей сути Data Science — это исследования и разработки. Каждый день специалисты берут данные, которые никогда ранее не анализировались, и ищут закономерности, не зная при этом точно, есть они там или нет. Дата-сайентисты взялись построить модели МО на основе данных, для которых еще может не быть новых трендов. Невозможно всегда получать успешный исход таких задач, потому что новые тенденции встречаются в какой-либо области очень редко. Однако в такой сфере, как разработка ПО, задачи обычно выполнимы (хотя это может занять больше времени и ресурсов, чем планировалось).

Важно понимать, как проваливаются проекты и что делать, если это произошло. Чем лучше вы понимаете причины, тем больше неудач сможете избежать в будущем. Неудачные проекты также могут дать представление о том, что оказалось успешным, если понять, какая часть все же сработала как надо. И возможно, вы сумеете поправить неудавшийся проект, превратив его во что-то полезное внутри организации.

В этой главе мы рассмотрим три темы: почему DS-проекты терпят неудачу, как учесть риски и что делать, если все идет не по плану. Мы обсудим три основные причины, по которым происходит большинство провалов, что делать с таким проектом и как справиться с эмоциями, которые вас могут одолевать.

13.1. Почему проваливаются DS-проекты

Кажется, что причин для провала не счесть. Помимо бюджета, технологий и задач, выполнение которых занимает гораздо больше времени, чем ожидалось, есть много других причин. В итоге все эти типы неудач можно разделить на несколько основных групп.

13.1.1. У вас не те данные, что вы хотели

Вы не можете изучить все возможные источники данных перед тем, как приступить к проекту. Абсолютно необходимо делать обоснованные предположения о том, что доступно, исходя из того, что вы знаете о компании. Когда работа над проектом начинается, вы часто обнаруживаете, что многие из ваших предположений не соответствуют действительности. Возможно, данных не существует, они не хранятся в удобном формате или в доступном для вас месте. Например, если вы проводите анализ, чтобы понять, как возраст клиента влияет на использование им программы лояльности, то можете узнать, что у клиентов просто не спрашивают возраст, когда те регистрируются в программе. Эта оплошность может очень быстро привести к завершению проекта.

Пример провала: анализ статуса программы лояльности

Директор отдела маркетинга крупной сети ресторанов хочет понять, тратят ли клиенты больше по мере повышения их статуса в программе лояльности компании. В программе есть серебряный, золотой и платиновый уровни, и директор хочет знать, покупает ли платиновый клиент столько же, сколько и на серебряном уровне.

Команда дата-сайентистов соглашается рассмотреть этот запрос, потому что задача кажется довольно простой, а данные о лояльности ранее не рассматривались. Но в процессе обнаружился неприятный сюрприз: устаревшая БД программ лояльности не отслеживает архивные данные, а фиксирует лишь то, на каком уровне клиенты находятся сейчас. Если клиент в настоящее время находится на платиновом уровне, невозможно узнать, когда он был на серебряном или золотом. Таким образом, провести анализ невозможно.

Команда рекомендует скорректировать систему, но для изменения архитектуры БД программы лояльности требуются миллионы долларов, а в компании нет на нее большого спроса, поэтому никаких изменений не вносится и от идеи проведения анализа отказываются.

Поскольку для того, чтобы приступить к работе, вам нужны данные, проблемы такого типа возникают в первую очередь. Обычной реакцией в таком случае является внутренний торг, в ходе которого вы пытаетесь залатать информацион-

ные дыры. Вы говорите что-то вроде: «Ну, у нас нет данных за десять лет, как мы хотели, но, возможно, для модели будет достаточно данных за год» и надеетесь на лучшее. Иногда этот подход может работать, но альтернативных решений не всегда бывает достаточно для выполнения задачи.

При работе над проектом у вас не всегда есть доступ к данным или даже полное понимание того, о чем идет речь (особая проблема в консалтинге, когда вы не получаете доступ к данным, пока работа для проекта не будет продана). Кроме того, данные могут быть предоставлены с критическим недостатком, который делает их бесполезными. Они могут существовать в таблице БД, но идентификаторы клиентов могут быть повреждены и непригодны для использования. Есть так много причин, по которым датасет может оказаться проблемным — а перед запуском проекта чрезвычайно сложно проверить их все! Поэтому часто происходит так, что DS-проекты обычно едва ли проходят стадию запуска.

Чем быстрее вы сможете получить доступ к датасету и изучить его, тем быстрее снизите риск неадекватных данных. Наилучший вариант, позволяющий избежать этой ошибки, — получить образцы данных перед запуском проекта. Если это невозможно, следующий вариант — разработать график реализации проекта с учетом вероятности, что данные будут плохими. Если на раннем этапе определить проект как «пойдет/не пойдет», когда стейкхолдеры соглашаются повторно оценить его осуществимость, то это снизит вероятность их удивления из-за возможной непригодности данных.

Если вы столкнулись с проблемой нехватки нормальных данных, то здесь не так много вариантов. Например, можно попытаться найти альтернативные источники. Возможно, у вас нет данных о приобретенных продуктах, но зато вы знаете, какой объем был произведен, и можете использовать эту информацию. Проблема обычно в том, что эти «заменители» существенно отличаются от оригинальных и могут привести к реальным проблемам при анализе.

Если вы не можете найти пригодную замену, иногда все, что вам остается, это начать отдельный проект по сбору более точных данных. Добавление инструментариев управления и телеметрии на веб-сайты и приложения, создание БД и другие способы могут помочь команде лучше справиться с поставленной задачей в будущем.

13.1.2. У данных нет сигнала

Предположим, что азартный игрок нанимает дата-сайентиста, надеясь использовать статистику для победы в игре в кости. Игрок бросает шестигранный кубик 10 000 раз и записывает броски; затем он платит аналитику за создание модели, которая прогнозирует следующий бросок кубика. Несмотря на то что специалист располагает огромным количеством данных, невозможно спрогнозировать, что

выпадет в следующий раз: можно только присвоить вероятность, равную 1/6 для каждой стороны (если кубик правильный). Хоть данных и много, в них нет сигнала о том, какая сторона выпадет в следующий раз.

Проблема отсутствия сигнала в данных встречается очень часто. Предположим, вы запускаете торговую онлайн-площадку и хотите создать модель, которая будет прогнозировать, какие клиенты будут заказывать продукты, в зависимости от их браузера, устройства и операционной системы. До запуска проекта нельзя узнать, действительно ли эти точки данных можно использовать для прогноза заказов; также нельзя определить, достаточно ли в данных сигнала, аналогично примеру с кубиком. Создание модели МО для прогнозирования означает тестировать данные, чтобы понять, есть ли в них сигнал, а его с большой вероятностью может и не быть. На самом деле во многих ситуациях удивительным будет *наличие* сигнала, а не его *отсутствие*.

Пример провала: обнаружение багов на веб-сайте с данными о продажах

У гипотетической компании, занимающейся электронной торговлей, есть проблема: на сайте постоянно появляются ошибки и баги. Хуже того, DevOps или команда инженеров ПО не всегда их обнаруживает. Однажды ошибку заметили маркетологи, обнаружив, что ежедневная выручка была слишком низкой. Если ошибку замечает маркетолог, а не DevOps или разработчики, дело плохо.

Команда дата-сайентистов намеревается использовать методы статистического контроля качества данных о продажах, чтобы получать предупреждения о чересчур низкой выручке, что, скорее всего, говорит о баге на сайте. Есть список дней, когда были обнаружены ошибки, и архивные данные о выручке. Использовать данные о продажах для прогнозирования ошибок кажется очевидным решением.

К сожалению, ряд причин, по которым выручка может меняться ежедневно, делает обнаружение багов практически невозможным. Она может быть низкой в конкретный день недели, из-за определенной даты, рекламных акций, глобальных событий или чего-то еще. Несмотря на то что когда-то маркетологам удалось обнаружить ошибку, этот факт не поддается обобщению, поскольку в данных не было сигнала о ней.

К сожалению, отсутствие сигнала в данных может привести к окончанию проекта. Если проект построен на попытке найти взаимосвязь и сделать на ее основании прогноз, а взаимосвязи нет, то прогноз невозможен. Анализ может не выявить ничего нового или интересного, или модель машинного обучения может не дать результатов, которые превзойдут случайные значения.

Если вам кажется, что вы не можете найти сигнал среди шума, есть несколько возможных выходов:

- *Пересмотрите проблему*. Можно попытаться переосмыслить проблему, чтобы понять, существует ли другой сигнал. Предположим, у вас есть набор статей

и вы пытаетесь спрогнозировать, какая из них будет наиболее релевантной для пользователя. Вы можете сформулировать проблему как задачу классификации, чтобы попытаться определить, какая статья в наборе является наиболее актуальной.

- *Измените источник данных*. Если вам кажется, что сигнала из данных нет, попробуйте изменить источник. Как и в предыдущем случае, связанном с отсутствием качественных данных, добавление нового источника иногда дает неожиданный сигнал. К сожалению, вы обычно начинаете работу с датасета, у которого были самые высокие шансы, поэтому маловероятно, что эта стратегия вас спасет.

Дата-сайентисты, застрявшие в подобной ситуации, обычно пытаются использовать более мощную модель для поиска сигнала. Если логистическая регрессия не может дать адекватный прогноз, пробуют случайную модель леса. Если и она не работает, пробуют нейронную сеть. Каждый метод требует больше времени и сил. Хотя эти методы могут быть полезны для получения более точных прогнозов, они не могут сделать что-то из ничего.

Чаще всего, если самый простой метод не может обнаружить какой-либо сигнал, более сложные методы тоже не помогут. Поэтому лучше начать с простых методов моделирования, чтобы проверить осуществимость проекта в принципе, а затем переходить к более сложным и трудоемким способам. Не следует тратить месяцы на построение все более сложных моделей в надежде, что, возможно, следующая станет спасательным кругом для проекта.

13.1.3. Проделанная работа оказалась не нужна

Независимо от того, насколько точны модель или анализ, важно, чтобы они приносили пользу стейкхолдеру. Результаты могут быть невероятно интересны дата-сайентистам, но не бизнесменам, которые их запрашивали. Модель MO может делать очень точные прогнозы, но если ее не развернуть и не использовать, она окажется бесполезной. Многие DS-проекты терпят неудачу даже после завершения работы.

Анализ данных, модель или сводная таблица — это продукт. Разработка и создание продукта — это практика, в которую многие люди вложили сотни лет коллективного мышления. Несмотря на все это, каждый год миллиарды долларов тратятся на создание продуктов, которые в итоге оказываются никому не нужны. С подобным сталкивались и New Coke, и Google Glass — некоторые высококлассные продукты не привлекают клиентов, не говоря уж о низкопробных. Подобно тому как Microsoft и Nokia могут приложить много усилий для создания Windows Phone, который никто не стал покупать, так и дата-сайентист может создавать продукты, которые не будут использоваться.

Пример провала: прогноз ценности кампании по организации и стимулированию сбыта

В розничной компании был запущен проект по созданию модели МО для прогнозирования окупаемости инвестиций (ROI) в будущие рекламные кампании. Команда специалистов по данным решила построить модель после того, как увидела, как отделы маркетинга и продаж бились над созданием таблиц Excel, которые прогнозировали бы общую ценность. Предположим, что, используя машинное обучение и моделирование на уровне клиента, группа дата-сайентистов создала модель на основе Python, которая более точно прогнозировала окупаемость инвестиций в кампании.

Позже команда выяснила, что единственная причина, по которой отдел маркетинга и продаж создал таблицы Excel с прогнозами окупаемости, заключалась в том, чтобы финансовый отдел их подписал. Специалисты финансового отдела отказались работать с чем либо, кроме Excel; Python был для них слишком сложной системой. Таким образом, ПО не использовалось, потому что команда дата-сайентистов не учла потребности клиента. Задача состояла не в получении максимально точного прогноза, а в том, чтобы получить прогноз, который убедил бы финансистов в целесообразности трат на рекламную кампанию.

Универсальный совет по созданию продуктов, которые понравятся покупателям, — проводить много времени, общаясь и взаимодействуя с заказчиками. Чем больше вы понимаете их потребности, желания и проблемы, тем больше шансов создать продукт, который им нужен. В изучении рынка и исследовании пользовательского опыта задействованы разные способы понимания клиента; в первом случае применяются опросы и фокус-группы, во втором — пользовательские истории, персоны покупателей и тестирования. Многие другие сферы бизнеса придумали собственные методы и используют их годами.

Несмотря на все старания, DS-проекты особенно подвержены провалам из-за непонимания потребностей клиентов. По какой-то причине специалистам по данным гораздо комфортнее смотреть на таблицы и графики, чем общаться с людьми. Многие такие проекты потерпели неудачу, потому что дата-сайентисты не приложили достаточно усилий, чтобы поговорить с клиентами и стейкхолдерами и понять суть их проблем. Вместо этого они занялись построением интересных моделей и исследованием данных. Фактически эта ситуация является одной из основных причин, по которой мы решили посвятить главу 12 работе со стейкхолдерами. Мы надеемся, что после ее прочтения вы стали лучше понимать, как выстраивать взаимоотношения с заинтересованными сторонами. Но если вы пропустили эту главу, возможно, лучше все же с ней ознакомиться.

Если вы оказались в ситуации, когда ваш продукт оказался без надобности, лучшее, что вы можете сделать, — это поговорить со своими клиентами. Это сделать никогда не поздно. Независимо от того, является ли этот человек бизнес-стейкхолдером или клиентом компании, общение всегда может быть полезно. Если

ваш продукт им не нужен, то могут ли они объяснить причину? Возможно, стоит добавить новые функции? А может, следует изменить анализ, добавив другой датасет. Пожалуй, получится улучшить модель МО, изменив формат выходных данных или скорость их выполнения. Вы никогда не узнаете о таких вещах, пока не поговорите с людьми.

Сюда также входит концепция минимально жизнеспособного продукта (minimally viable product, MVP), которая широко используется при разработке ПО. Идея состоит в том, что чем быстрее вы сможете заставить продукт работать и вывести его на рынок, тем быстрее узнаете, что работает, а что нет, и сможете учесть это в следующий раз. Чем скорее начнет работать какая-либо модель или будет выполнен анализ, тем раньше вы сможете показать их клиентам или стейкхолдерам и получить отзывы. Если вы потратите месяцы на итерацию модели, то не сможете получить эту обратную связь.

Чем лучше вы понимаете клиентов на протяжении всего процесса разработки, тем меньше вероятность того, что вы потерпите неудачу из-за того, что ваш продукт не захотят использовать. Но если это все же произойдет, лучший способ двигаться дальше — начать диалог, чтобы попытаться найти решение.

13.2. Управление риском

Одни проекты более рискованные, чем другие. Использование данных, с которыми команда уже работала, и создание стандартной сводной таблицы обычным способом, скорее всего, приведут к желаемому результату. Но поиск нового датасета, построение на его основе модели МО, которая будет работать в режиме реального времени, и создание приятного пользовательского интерфейса — это более рискованный проект. Как дата-сайентист вы можете в какой-то мере контролировать степень риска.

Один из важных факторов риска — это количество проектов, над которыми вы работаете одновременно. Если у вас всего один рискованный проект и он провалился, то справиться с этой неудачей может быть довольно сложно. Но работая над несколькими проектами одновременно, вы сможете снизить риск: если провалится один из проектов, у вас есть другие. Если один проект представляет собой чрезвычайно сложную модель МО, которая имеет ограниченные шансы на успех, вы можете параллельно работать над более простыми сводными таблицами и отчетностью; тогда, если проект машинного обучения не удастся, стейкхолдеры все равно останутся довольны отчетами.

Наличие нескольких проектов также может быть выгодно с точки зрения загрузки. В DS-проектах часто возникают паузы из-за ожидания данных, ответа от стейкхолддера или подбора моделей. Если вы по какой-то причине застряли

с одним проектом, у вас будет возможность переключиться на другой. Это может даже помочь с ментальными блоками; если вы завязли в проблеме, отличный способ освежить мысли — отвлечься.

Еще один способ снизить риск — заложить в проект раннюю остановку. В идеале потенциально рискованный проект должен разрабатываться с расчетом на то, что если к определенному моменту с ним ничего не выйдет, то работа будет прервана. Например, если вам поручают проект и при этом неясно, есть ли необходимые для него данные, то границы проекта можно определить так: если через месяц поиска подходящие данные найти не удастся, то проект будет сочтен нереализуемым и аннулируется. Если учесть риски проекта на ранней стадии, то его завершение будет менее неожиданным и затратным.

В некотором смысле раннее завершение проекта говорит о том, что Data Science — это исследования и разработки. Поскольку в этой сфере так много неизвестного, есть смысл предусмотреть возможность того, что по мере получения большего количества информации в ходе исследования идея может не сработать.

Несмотря на необходимость минимизировать риски портфеля проектов, не стоит полностью их исключать. Data Science вообще строится на рисках: практически любой достаточно интересный проект будет полон неопределенности и неизвестности. Это может быть из-за того, что никто в компании раньше не использовал новый датасет, не пробовал конкретную методологию или же стейкхолдер работает в департаменте, который никогда раньше не прибегал к помощи Data Science. Огромный вклад в эту сферу внесли люди, пробующие что-то новое, и если вы как специалист по данным пытаетесь избегать потенциально рискованных проектов, вы также лишаете себя шанса на большой успех.

В этой главе перечислены разные причины, по которым DS-проект может провалиться, но нежелание рисковать может в конечном итоге вылиться в провал целой группы дата-сайентистов. Представьте себе команду, которая придумывает новые идеи и отчеты по проектам, считает их успешными, а затем начинает «буксовать», обновляя только предыдущую работу. В таком случае сотрудники упустят потенциальные новые возможности.

13.3. Что делать, если проекты терпят неудачу

Если ваш DS-проект провалился, это не значит, что все время, которое вы над ним работали, потрачено зря. В разделе 13.2 мы изложили некоторые возможные варианты для улучшения проекта. Но даже если это не поможет, все же есть шаги, которые вы можете предпринять, чтобы использовать то, что осталось от проекта, по максимуму. В следующих разделах мы расскажем о нескольких стратегиях, которые помогут справиться с эмоциями в случае неудачного проекта.

13.3.1. Что делать с проектом

Несмотря на то что проект мог провалиться, в нем наверняка остались полезные штуки, которые еще пригодятся. Следующие шаги могут помочь вам сохранить многие из них.

ДОКУМЕНТАЦИЯ

Первое, что нужно сделать с неудачным проектом, — это оценить, чему он может научить вас. Некоторые важные вопросы, которые следует задать себе и команде:

- *Почему не получилось*? Этот вопрос кажется очевидным, но часто бывает так, что вы не можете понять, почему проект провалился, пока не вернетесь назад и не посмотрите на картину в целом. Поговорив со всеми участниками проекта, вы сможете лучше понять, что пошло не так. Компания Etsy популяризировала концепцию *безобвинительного «разбора полетов»* — обсуждения, проводимого после неудачи, в ходе которого команда может диагностировать проблему, не обвиняя человека. Если вы думаете, что проблема связана с работой команды (а не ошибками конкретного человека), вы с большей вероятностью найдете решение. Не опасаясь последствий, люди охотнее говорят открыто о произошедшем.
- *Что можно было сделать, чтобы предотвратить провал*? Зная причины неудачи, вы сможете понять, как избежать подобных ситуаций в будущем. Например, если для работы проекта было недостаточно данных, то провал можно было предотвратить с помощью более длительной исследовательской фазы. Подобные уроки помогают команде расти и развиваться.
- *Что вы узнали о данных и о задаче*? Даже если проект потерпел неудачу, вы часто узнаете то, что пригодится в будущем. Возможно, в данных не было сигнала, но, чтобы понять это, вам все равно пришлось присоединить несколько новых датасетов; теперь вам будет легче выполнять аналогичные действия в других проектах. Подобные вопросы помогут провести мозговой штурм, понять, что можно спасти в этом проекте, и придумать альтернативные идеи.

Если вы проведете совещание, на котором команда проработает эти вопросы, а затем сделаете выводы общедоступными, то неудачный проект принесет гораздо больше пользы.

РАССМОТРИТЕ ВОЗМОЖНОСТЬ ИЗМЕНЕНИЯ ПРОЕКТА

Несмотря на то что сам проект мог потерпеть неудачу, бывают способы превратить его во что-то полезное. Если вы пытаетесь создать ПО для обнаружения аномалий в доходах компании, а у вас ничего не получается, вы все равно можете использовать ту же модель в качестве довольно неплохого инструмента про-

гнозирования. Целые компании были построены на неудачных идеях, которые превратились в нечто успешное.

Изменение продукта требует активного взаимодействия со стейкхолдерами и клиентами. По сути вы снова оказываетесь в самом начале разработки, пытаясь найти хорошее применение своей работе. Поговорив со стейкхолдерами и клиентами, вы сможете понять их проблемы и увидеть, пригодится ли ваша работа для чего-то другого.

ЗАВЕРШЕНИЕ ПРОЕКТА (ПОСПЕШНОЕ)

Если у вас не получается изменить проект, лучшее, что вы можете сделать, — это завершить его. Окончательно отказываясь от проекта, вы даете себе и своей команде возможность перейти к новой, более перспективной работе. Дата-сайентист может хотеть развивать проект вечно в надежде, что когда-нибудь он заработает. (Есть тысячи алгоритмов; в конце концов, один из них сработает, ведь так?) Но если что-то не работает и вы с этим застряли, в конечном итоге вы только напрасно потратите силы. Кроме того, работать над одним и тем же проектом до скончания веков не особо приятно! Несмотря на то что закрывать проект всегда сложно (ведь для этого нужно признать, что не нужно больше тратить на него силы), в конечном итоге оно того стоит.

ОБЩАЙТЕСЬ СО СТЕЙКХОЛДЕРАМИ

Дата-сайентист должен поддерживать связь со стейкхолдерами на протяжении всего проекта (см. главу 12); количество диалогов должно быть удвоено, если проект близок к провалу. Всегда есть соблазн скрыть риски и проблемы, дабы не разочаровывать стейкхолдера, но ситуация, когда он все же узнает о провале, может иметь катастрофические последствия для карьеры. Сообщая о текущих проблемах или о том, что проект больше не может продвигаться, вы обеспечиваете прозрачность своей работы и внушаете доверие. Помогая стейкхолдеру понять состояние проекта, вы можете определить следующие шаги совместными усилиями.

Если вы не знаете, как сообщить стейкхолдеру о проблемах, обратитесь к своему руководителю за помощью. Он может либо обсудить с вами, как лучше преподнести информацию, либо взять эту ответственность на себя. У каждого свои предпочтения в том, как лучше узнавать о проблемах: кто-то любит получать таблицы, где вопросы выделены зеленым/желтым/красным цветом, а кто-то предпочитает обсудить все за чашкой кофе. Ваш руководитель или другие члены команды должны знать, что работает лучше всего.

Как дата-сайентист вы часто волнуетесь, сообщая о провале проекта, чувствуете себя очень уязвимым эмоционально и думаете, что находитесь в слабой позиции. Несмотря на то что иногда подобные новости воспринимаются плохо,

обычно люди готовы помочь с проблемами и принятием решения о дальнейших действиях. Сообщив о провале проекта, вы можете почувствовать облегчение.

13.3.2. Как справиться с негативными эмоциями

Забудьте ненадолго о проекте и компании: нужно думать и о собственном здоровье. Провал с точки зрения эмоций — это сложно! И это самое ужасное! Если вы не уделите достаточно внимания своим эмоциям, мысли о неудаче могут преследовать вас еще очень долго после закрытия проекта. Если вы внимательно отнесетесь к своей реакции на провал и к тому, что и как вы будете о нем говорить, то сможете настроить себя на успех в долгосрочной перспективе.

Естественный внутренний монолог в конце провалившегося проекта звучит так: «Будь я хорошим специалистом, проект бы не провалился». Это заблуждение. Большинство проектов не получается, потому что сфера Data Science по своей сути основана на опробовании вещей, которые никогда не сработают. Большинство крутых специалистов по данным либо имеют за плечами проекты, которые не увенчались успехом, либо даже руководили ими. Возлагая вину за провал на себя и на недостатки Data Science, вы перекладываете на себя всю тяжесть проекта. Но, как мы уже говорили, есть много причин, по которым проекты терпят неудачу, и очень редко они кроются в компетентности дата-сайентиста. Беспокоиться о том, что проект провалился из-за вас, — распространенная проблема, но все это находится исключительно в вашей голове и никак не отражает реальность.

Если вы позволите себе ошибаться и примете тот факт, что неудача — не признак слабости, у вас будет больше возможностей извлечь уроки из этого опыта. Уверенность в себе и своих способностях позволяет легче переживать провалы и их причины. При этом способность быть уверенным в себе и признавать неудачу требует времени, терпения и практики, поэтому не удивляйтесь, если это сложно дается. Так и должно быть!

Лучшее, что вы можете сделать для себя, если проект идет ко дну, — понимать, что неудача не отражает ваши навыки. Проекты терпят крах по причинам, не зависящим от вас, а вы можете идти дальше. Чем больше вы будете контролировать свои эмоции, тем легче будет принять ситуацию.

Мы закончим эту главу метафорой. Новички и джуниоры воспринимают своих коллег-профессионалов как архитекторов зданий. Начинающий архитектор может проектировать простые дома, а опытный — строить небоскребы, но если один из них ошибется и спроектированное им здание рухнет, это приведет к краху карьеры. Аналогичным образом можно рассматривать дата-сайентиста: он строит все более сложные и замысловатые модели, но если одна из них окажется неудачной, то вся карьера окажется под угрозой. Мы надеемся, что, прочитав эту главу, вы поймете, *что на самом деле все немного иначе.*

Есть другая, более подходящая метафора: дата-сайентист подобен охотнику за сокровищами (рис. 13.1). Охотник за сокровищами отправляется на поиски кладов, и, если ему повезет, он найдет их! Начинающий охотник может искать стандартные безделушки, а опытный находит самые легендарные сокровища. Специалист по данным больше похож на охотника за сокровищами; он ищет эффективные модели, и порой эти модели и анализ работают! Несмотря на то что синьор может работать над более сложными или заковыристыми проектами, он тоже может потерпеть неудачу, и это лишь часть работы.

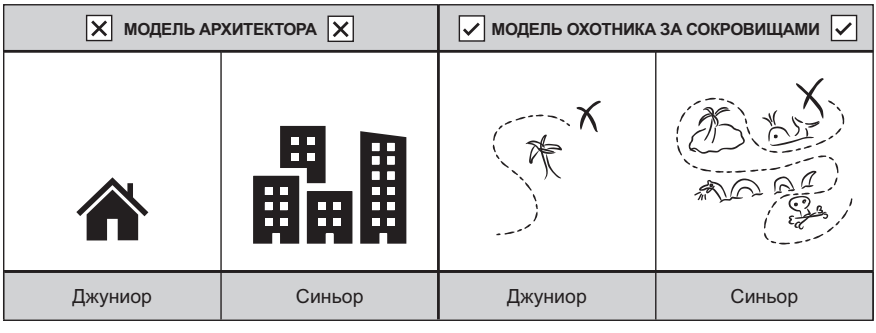


Рис. 13.1. Две метафоры: архитектор и охотник за сокровищами

13.4. Интервью с Мишель Кейм, руководителем отдела Data Science и машинного обучения Pluralsight

Мишель Кейм (Michelle Keim) руководит группой по Data Science и МО в Pluralsight, платформе обучения корпоративным технологиям, миссией которой является демократизация технологических навыков. До этого она работала и возглавляла DS-команды в различных компаниях, включая Boeing, T-Mobile и Bridgepoint Education, поэтому ей хорошо известно, по каким причинам проекты в Data Science могут терпеть неудачу и как с этим справляться.

Расскажите о своем опыте провалившегося проекта.

Меня пригласили на должность руководителя проекта по построению набора моделей удержания клиентов. Я думала, что поговорила со всеми нужными стейкхолдерами и поняла потребности бизнеса, разобралась, как работает команда и зачем нужны модели. Мы разработали их, но вскоре узнали, что модели остались невостребованными. Проблема заключалась в том, что мы не встретились со специалистами по обслуживанию клиентов, которые фактически использовали

бы их. Я разговаривала только с руководителями. Мы составили перечень вероятностей ухода клиента, но специалисты по обслуживанию не знали, что с ним делать. Им нужно было понимать, что делать, когда возникает риск ухода клиента, а это проблема совершенно другого рода, чем та, которую мы пытались решить. Самый значимый урок, который я извлекла, — действительно важно погрузиться в задачу и определить, почему она появилась. Какую задачу решают люди, которые будут использовать результат?

Есть ли какие-то индикаторы риска, которые вы можете заметить до начала проекта?

Я думаю, что отчасти это инстинкт, который развивается с опытом. Чем больше вы видите случаев, когда что-то идет не так, и чем чаще вы пользуетесь возможностью извлечь уроки из неудач, тем лучше вы видите, где искать тревожные признаки. Ключ к решению в том, чтобы сократить цикл и обнаружить тревожные сигналы раньше. Нужно как можно чаще получать обратную связь.

Дата-сайентисты обычно с головой погружаются в работу и забывают обо всем остальном. Очень важно иметь представление не только о том, чего вы хотите достичь к концу дня, но и о том, каким должен быть результат на разных этапах. Благодаря этому вы сможете оценивать текущие результаты, получать обратную связь и при необходимости изменять направление. Если вы что-то упустили или неправильно поняли, контрольные точки позволяют быстро это найти и исправить. В противном случае вы можете узнать об этом в конце, и работу придется переделывать.

Как разные компании решают вопрос провалившегося проекта?

Это тесно связано с корпоративной культурой. Я бы посоветовала при поиске работы попытаться выяснить, есть ли в компании культура обучения и постоянной обратной связи. На интервью у вас есть возможность спросить: «Вы учитесь чему-нибудь опытным путем? Как у вас появилась такая возможность? Если бы я занял эту должность, как бы я получал обратную связь о своей работе? Есть определенные правила или мне нужно будет искать способы самостоятельно?» Ответы сотрудников на эти вопросы расскажут о многом.

Если вы уже работаете в компании, можете задать себе несколько вопросов, чтобы понять, здоровая ли в компании атмосфера. Есть ли после завершения проекта возможность остановиться и проанализировать сделанное? Сможете ли вы учиться на ошибках по завершении проекта? Видите ли вы, что руководство на различных уровнях открыто для общения и берет на себя ответственность за неудачи? В условиях отсутствия развитой культуры вы будете чувствовать страх и видеть поведение, направленное на удовлетворение корыстных целей, а не на выполнение общей миссии: это будет вас угнетать.

Как понять, что проект, над которым вы работаете, идет ко дну?

Вы не сможете понять, что терпите неудачу, если в самом начале не определили, что такое успех. Каких целей вы пытаетесь достичь и как выглядят контрольные точки на пути к успеху? Если вы этого не знаете, то можете лишь гадать, хорошо ли продвигается проект. Для достижения успеха вы должны убедиться, что в достаточной мере сотрудничаете со стейкхолдерами, чтобы получить четкий ответ на эти вопросы. Вы должны знать, зачем выполняете конкретный проект и какую задачу пытаетесь решить, иначе не сможете оценить, целесообразны ли ваши предложения и верен ли ваш подход. Одна из обязанностей дата-сайентиста состоит в том, чтобы, поделившись своим опытом, помочь сформулировать задачу и показатели успеха.

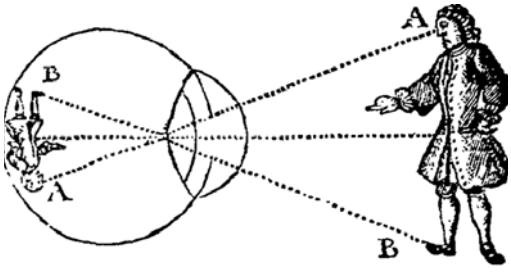
Как вы преодолеваете страх неудачи?

Нужно помнить, что ошибки необходимы: если все будет идти идеально, то ничему не научитесь. Как иначе вы будете развиваться? Этот опыт нужен, поскольку нет альтернативы тому, чтобы справиться с неудачей и перестать переживать о ней. Провал может быть болезненным и привести вас к вопросу: «Черт возьми, что же мне делать?» Но после того, как вы успокоитесь, сделайте выводы и превратите их в навык. Стойкость к неудачам со временем превратится в уверенность в собственных силах. Если вы знаете, что может пойти не так, в следующий раз вам будет легче. И если вы будете получать обратную связь достаточно часто, то сможете обнаружить ошибки прежде, чем они приведут к разрушительным последствиям. Никто не ждет от вас совершенства. От вас ждут, что вы будете честно говорить о том, чего не знаете, и продолжите учиться, задавая вопросы и учитывая обратную связь.

Итоги

- Причиной провала проекта могут стать неадекватные данные, отсутствие сигнала или отказ клиента использовать результат.
- После того как проект провалился, выясните причины и подумайте о внесении изменений или об отказе от проекта.
- Провал проекта никак не характеризует квалификацию дата-сайентиста.
- Дата-сайентист не несет единоличной ответственности за провал проекта.

14



Вступление в сообщество Data Science

В этой главе

- Как расширить портфолио проектов и постов в блогах.
- Как найти и извлечь максимальную выгоду из конференций.
- Как выступать на митапах или конференциях.
- Как вносить свой вклад в открытый исходный код.

Вам может казаться, что упорный труд на должности в Data Science — это единственная возможность продвинуться по карьерной лестнице. Но есть много других способов развить свои навыки, особенно за счет взаимодействия с сообществом дата-сайентистов. Внерабочая деятельность — например, выступление с докладами или участие в разработке ПО с открытым исходным кодом — может быть чрезвычайно полезной для карьеры.

В этой главе мы рассмотрим четыре способа вступления в сообщество: расширение портфолио, посещение конференций, выступление с докладами и участие в разработке открытого исходного кода. Мы предлагаем четыре варианта, чтобы у вас был выбор; у очень немногих людей есть время и силы на все сразу. Хотя эти дела отнимают время и являются дополнением к повседневной работе, это не значит, что они должны полностью поглотить вашу жизнь. В этой главе мы дадим советы по эффективному использованию своего времени с помощью таких так-

тик, как многократное использование докладов, превращение публикации в блоге в доклад и написание статьи о своем первом вкладе в открытый исходный код.

Рис. 14.1. Некоторые из способов вступить в сообщество, описанные в этой главе

Все это может быть полезным и чрезвычайно интересным, но вам вовсе не обязательно этим заниматься, чтобы сделать карьеру в Data Science. Многие специалисты по данным, в том числе руководители в ведущих компаниях, не делают ничего из предложенного. Но мы, авторы этой книги, считаем, что сообщество единомышленников много раз помогало нам в карьере, в том числе в получении офферов и в повышении. В общественной деятельности потраченное время окупается вдвойне.

Мы видим четыре основных преимущества в присоединении к более широкому сообществу дата-сайентистов (рис. 14.1):

- *Приобретение навыков*. Взаимодействуя с сообществом, вы узнаете о новых методах, которыми бы не воспользовались, если бы полагались только на то, чем занимаетесь на работе ежедневно. Создание проекта с открытым исходным кодом — это деятельность, которая самым непосредственным образом развивает технические навыки, потому что вы будете писать код для сторонних пользователей и работать совместно с другими. У каждого вида деятельности есть свои преимущества. Ведение блога — отличный способ выявить пробелы в своих знаниях и получить обратную связь. Лекции помогают отточить навыки презентации, благодаря которым вы однажды сможете убедить стейкхолдера в том, что вам нужно финансирование или что он должен поддержать ваш проект. Информация, полученная на конференции, может сдвинуть с мертвой точки важный проект и сэкономить часы работы.
- *Расширение сети знакомств*. Связь с сообществом — отличный способ найти группу людей по интересам, которые понимают ваши трудности. Даже несмотря на друзей среди коллег, вам может не хватать опыта в определенной нише; совет, данный кем-то из членов сообщества, поможет заполнить этот пробел. Вы также можете узнать, какие условия предлагают другие компании.
- *Получение возможностей*. Чем больше участия вы принимаете в жизни сообщества, тем чаще вас будут просить помочь с проектом, выступить с докладами или записаться в подкасте. Вы даже можете получить следующую работу,

благодаря тому что кто-то найдет вас по работе, которую вы выложили в интернете, или встретит вас на конференции. Это большой цикл положительной обратной связи: выступления приводят к большему количеству выступлений, а проекты приводят к большему количеству проектов. Эти возможности могут быть информативными, интересными и увлекательными.

- *Отдача*. Это не столько прямая выгода для вас самих, сколько вклад в развитие общества. Если вы спросите наставника, как вы можете отблагодарить его за поддержку, многие скажут: «Продолжи цепочку. Помоги другим и стань для них наставником». Принадлежность к сообществу может сделать работу в Data Science намного интереснее. Выполняя задачи, которые помогают другим, вы будете чувствовать свою ценность, словно работаете не только ради зарплаты.

14.1. Расширение портфолио

То, что у вас есть работа, не означает, что можно забыть все прекрасные привычки, которые у вас выработались ради ее получения. В главе 4 вы узнали, как писать в блогах и создавать портфолио. Наличие полноценной работы не означает, что нужно все это забросить. При этом блог или сторонний проект не должен быть обременительным; в этой главе мы расскажем, как извлечь максимум при минимуме усилий.

14.1.1. Больше публикаций

Мы надеемся, что вы узнаете много нового на должности дата-сайентиста. Как можно оптимизировать SQL-запросы к таблице с 30 миллиардами строк? Как эффективно работать с маркетологами? Какие стратегии использовать для навигации среди сотен таблиц?

Если вы работаете в компании, где есть другие специалисты по данным, то будете учиться непосредственно у них, читая код коллег или занимаясь парным программированием. Мы рекомендуем вам все записывать, так как вы будете получать много новой информации и вряд ли сможете все это вспомнить через несколько месяцев. А почему бы не поделиться своими записями с классом (в данном случае со всем интернетом)? От этого выиграют не только анонимные читатели: публикация статей — отличный способ закрепить знания. Вы можете вернуться к написанному даже через несколько лет.

Если вы последовали нашему совету из главы 4, у вас уже должен быть блог с несколькими статьями. Если вы этого не сделали, но хотите начать, мы рекомендуем вернуться к этой главе и следовать ее указаниям. Все, что мы там написали, по-прежнему актуально, те же стратегии, благодаря которым вы создали успешные публикации при поиске первой работы в Data Science, работают даже

после полноценного трудоустройства. Единственное серьезное изменение в том, что, если вы пишете о проектах, которые выполняли на работе (за исключением общих приобретенных навыков программирования, статистики или управления персоналом), необходимо убедиться, что вы не раскрываете конфиденциальную или служебную информацию и соблюдаете любые другие правила, установленные компанией для личных блогов сотрудников (например, сначала пост должен быть опубликован PR-отделом).

Если вы не хотите вести личную платформу, проверьте, есть ли у компании технический блог. Даже если предыдущие публикации были ориентированы на инженерию, вы можете написать пост о данных. Одобрение публикации может затянуться, но вы получите бонус в виде возможности записывать свои мысли в рабочее время. Даже если компания не ведет блог, у нее должны быть какая-нибудь внутренняя документация и программа обучения. Если вам пришлось чему-то научиться, задавая вопросы коллегам или просматривая устаревшие инструкции, создайте или обновите ресурс с четкими инструкциями для будущих новых сотрудников. Если этот вопрос пригодится кому-то за пределами компании (не только в качестве вашего собственного ПО для внутреннего пользования или для описания данных), то позже из него можно сделать статью в блоге или доклад.

14.1.2. Больше проектов

DS-проекты (которые мы также рассмотрели в главе 4) — это те, в которых вы выбираете или создаете датасет и анализируете его, чтобы ответить на вопрос. Можно использовать Twitter API, чтобы проанализировать методом сеток тех пользователей, которые пишут твиты о конференциях по Data Science. В некоторых случаях проект даже не обязательно должен представлять собой анализ; может быть, вы решите похвастаться своими инженерными навыками, создав Slack-бота, чтобы пользователи могли начислять друг другу «баллы», отслеживая итоги в созданной вами БД.

Работать с проектами может быть сложнее, чем писать в блог. В зависимости от отрасли компания может даже поощрять вас, если вы будете в общих чертах описывать свою работу. Даже если и нет, все равно можно писать нетехнические посты о том, как общаться с бизнес-стейкхолдерами или рассказывать о своем опыте на рынке труда. Учтите, что очень немногие компании делятся данными публично, поэтому даже если бы вы и могли выложить свой код потрясающего анализа, это было бы бессмысленно, ведь данными или результатами поделиться нельзя. Если вы хотите поделиться своим анализом, то придется заниматься им как сторонним проектом строго в свободное от работы время.

С другой стороны, периодически все же стоит брать сторонние проекты. Например, если вы захотите сменить работу, компании могут попросить предоставить какой-нибудь образец анализа данных. Если вы проработали в компании несколько лет, то не стоит показывать проект с курсов или для MOOC, ведь вам нужно продемонстрировать, как развивались ваши навыки. При этом принципы выбора темы и написания хорошего анализа те же, что и описанные в главе 4.

Хорошая новость в том, что проект не должен отнимать много времени. Дэвид Робинсон, интервью с которым представлено в главе 4, еженедельно создает скринкаст, в котором записывает, как он анализирует абсолютно новый датасет (из проекта Tidy Tuesday https://github.com/rfordatascience/tidytuesday/). На этот анализ у него уходит около часа, поскольку он не занимается подготовкой данных, при этом загруженный на GitHub код может служить примером проекта анализа. Для того чтобы так быстро и качественно управиться, дата-сайентист должен быть довольно опытным, но любой может попытаться ограничить себя по времени выполнения и сосредоточиться на обмене результатами (а не тратить 14 часов на проект, который никто никогда не увидит).

14.2. Посещение конференций

Быть частью сообщества означает иногда выбираться из дома. По большей части имеется в виду посещение конференций, где люди, работающие (или желающие работать) в аналогичных сферах, собираются вместе, чтобы поделиться опытом.

Конференции обычно представляют собой ежегодные мероприятия, которые проходят по всей стране и по всему миру. В области Data Science проводится много конференций: Strata, rstudio::conf, PyData, EARL и Open Data Science Conference — лишь некоторые из наиболее крупных, и у большинства из них есть региональные представительства. Вам также могут быть интересны технологические конференции на более общие темы, которые пересекаются с Data Science: Write/Speak/Code, PyCon, Grace Hopper, and SciPy.

Длятся конференции, как правило, от двух до четырех дней, а программа расписана на целый день — с раннего утра до вечера, после чего проводятся коллективные мероприятия. Они могут быть «однодорожечными» (одновременно происходит только одно выступление) или «многодорожечными» (выступает несколько участников сразу), но в конференциях всегда участвует несколько спикеров. Участие в подобном мероприятии может быть очень дорогим — обычно от $300 до $700 в день только за билет. На некоторых конференциях за дополнительную плату (обычно около $750 в день) также проводятся семинары, которые могут занимать от нескольких часов до двух дней.

Одна из причин, по которой мы указываем приблизительные цены, заключается в том, что есть много способов получить скидку. Если вы относитесь к недостаточно представленной социальной группе, поищите стипендии или промокоды, которые предлагаются всем таким людям, например R-Ladies или PyLadies. Если вы работаете в некоммерческой организации или в научном сообществе, то также можете платить меньше; кроме того, многие крупные конференции предоставляют скидки при ранней покупке билета. Еще один отличный способ снизить расходы — выступить с докладом, поскольку спикерам конференции должны предоставлять бесплатный вход. Наконец, некоторые конференции выдают стипендии, которые покроют стоимость билета и, возможно, все расходы, включая транспорт и проживание. Можете попробовать подать заявку.

Раз уж речь зашла о потенциально высокой цене, то зачем вам тратить время и деньги (или деньги работодателя) на участие, особенно если организаторы конференции записывают и размещают выступления онлайн? Здесь мы возвращаемся к одному из основных преимуществ, о котором говорили в начале главы: формирование сети контактов. Нетворкинг может иметь негативный подтекст: некий человек ходит по залу, обменивается рукопожатиями в надежде встретить кого-то важного и что-то от него получить. Но в лучшем своем проявлении нетворкинг — это поиск сообщества людей, которые вас поддерживают. Эта поддержка может быть очень ощутимой, например когда вас знакомят с сотрудником компании, на вакансию которой вы откликались, Или же это может быть нечто неосязаемое, как, например, то чувство, когда вы наконец-то оказались в зале, полном технических специалистов, большинство из которых — женщины.

Нетворкинг лучше всего работает в долгосрочной перспективе, поэтому, даже если вы думаете, что сейчас вам помощь не нужна, было бы неплохо заложить фундамент до того, как начать искать новую работу или партнера для проекта по написанию открытого исходного кода.

ДРЕСС-КОД Те, кто впервые идут на мероприятие, часто задаются вопросом, что надеть. В целом для конференции подойдет стиль кэжуал. Если речь идет о конкретном мероприятии, поищите фотографии с него в Twitter или на веб-сайте. Однако имейте в виду, что выступающие могут быть одеты более формально, чем зрители; тот факт, что все спикеры придерживаются стиля бизнес-кэжуал, не означает, что их слушатели тоже должны это делать. Если вы действительно не знаете, как одеться, возьмите с собой универсальные вещи, например повседневное платье или футболку поло и темные джинсы. Организаторы конференций редко впадают в крайности, устанавливая дресс-код в виде костюма или футболки и шортов. Кроме того, обычно наблюдается некоторое разнообразие стилей, поэтому вы вряд ли будете слишком выделяться.

Как при таком большом разнообразии конференций выбрать те, которые стоит посетить? Вот несколько направлений, которые следует учитывать:

- *Академические*. Некоторые конференции, такие как useR!, NeurIPS и JSM, собирают большое количество участников из академических кругов или специалистов, занимающихся серьезной исследовательской работой. В крайнем случае бывают конференции, на которых практически каждый является аспирантом или профессором. Если вы работаете в корпорации, то вряд ли услышите много выступлений на интересующие вас прикладные темы. Несмотря на то что докладчиками могут быть и представители индустрии, их выступления могут быть посвящены передовым алгоритмам машинного обучения, а эта тема актуальна, только если вы работаете в гигантской компании онлайн-торговли.
- *Масштаб*. Конференция может насчитывать от 150 до десятков тысяч человек. Мы рекомендуем начать с небольшой или средней конференции — от 200 до 1500 человек. Меньший размер означает, что навигация будет менее сложной и вы с большей вероятностью встретите одних и тех же людей несколько раз, что приведет к более прочным связям.
- *Компании, нанимающие сотрудников*. С другой стороны, вы можете пойти на большую конференцию, потому что ищете работу. Хотя такие компании можно встретить на любой конференции, некоторые более крупные из них проводят ярмарки вакансий, где работодатели отдельно платят за установку специального стенда и общение с потенциальными сотрудниками.
- *Уровень докладов*. Большинство конференций обычно нацелены на людей, которые работают или учатся в этой области. Если вы не знаете R, то вряд ли много вынесете с rstudio::conf — конференции, которая проводится компанией, занимающейся основной интерактивной средой разработки (IDE) для R. Доклады обычно нацелены на середнячков в плане общих знаний, но ожидаемый уровень специальных навыков может различаться. Например, на конференции rstudio::conf можно услышать презентацию о пакете для временны́х рядов. Вы должны немного знать R, чтобы понимать ее, но докладчик не ждет от вас большого опыта работы с временны́ми рядами. Или доклад об онлайн-экспериментах может стать введением в тему о том, как качественные исследования могут дополнять количественные методы.
- *Разнообразие и инклюзивность*. К сожалению, не все организаторы думают о комфорте всех ее участников. Если вы видите, что все 45 докладчиков — мужчины, то можно с уверенностью предположить, что аудитория слушателей тоже будет мужской. Помимо списка выступающих посмотрите на сайте, есть ли какие-либо правила поведения. Если вам нужны определенные условия,

например место для инвалидной коляски, найдите адрес электронной почты организаторов и отправьте письмо с запросом.

- *Специализация*. Конференции, так же как и Data Science, имеют множество специализаций. Если вам нужны конкретный язык или предметная область, наверняка найдется подходящее мероприятие.

Когда вы определились, какой именно тип конференции вам интересен, для начала поищите обзоры на них. Если вы лично не знаете никого, кто уже бывал там, задайте вопрос в Twitter или LinkedIn. Также посмотрите планируемое или, если его еще нет, прошлогоднее расписание. Если есть записи выступлений, посмотрите несколько из них. Следует убедиться в том, что ваши инвестиции окупятся. К сожалению, на некоторых конференциях не так много хороших спикеров.

Посещение конференций дает вам возможность продвигаться вверх по карьерной лестнице; кроме того, ваш работодатель от них тоже выигрывает, поскольку там вы можете рассказывать о своей компании и узнавать то, что сделает вашу работу эффективнее. Поэтому можно убедить работодателя оплатить ваше участие полностью или частично. У некоторых компаний есть официальный бюджет на конференции или обучение, и это здорово, поскольку деньги выделяются для ваших целей, но, если вы захотите выйти за рамки финансирования, вам будет сложно убедить компанию сделать исключение.

Помимо затрат на участие стоит учитывать фактор времени. Большинство конференций проводятся в течение недели (по крайней мере, частично), поэтому вам придется отпроситься с работы на один или два дня. Вы же не хотите, чтобы эти дни вычли из вашего отпуска, верно? Вы должны убедить руководителя, что вам стоит провести день именно там, а не заниматься обычной работой. У некоторых компаний есть политика в отношении количества дней, которые вы можете взять для участия в таких мероприятиях. В технологических компаниях вполне нормально посетить хотя бы одну конференцию, но в других отраслях такой системы может не быть.

Если вы хотите убедить руководителя в том, чтобы он отпустил вас, сделайте акцент на следующем:

- *Рекрутинг*. Наем дата-сайентиста стоит тысячи или даже десятки тысяч долларов. Одна из самых сложных задач — вообще привлечь хороших кандидатов. Крупные технологические компании, такие как Google и Amazon, а также известные стартапы обычно не испытывают недостатка в отличных заявках, но большинство компаний не пользуются такой популярностью. Если вы встречаетесь с людьми на конференциях, то рассказываете о своей компании. Эта реклама усиливается еще больше, если вы говорите то, о чем мы расскажем в разделе 14.3.
- *Знания*. Ваш руководитель захочет узнать, что такого вы сможете делать после конференции, чего не умели раньше. Еще лучше, если вы поделитесь полу-

ченными знаниями с командой, написав статью (которая также может стать публикацией в блоге!) или презентацию. Начните просматривать расписание конференции и сообщайте руководителю о докладах, информацию из которых вы сможете применять в работе. Имейте в виду, что кроме презентаций на конференциях также есть неформальное общение. Вы можете встретить человека, который знает, как решить проблему, с которой вы столкнулись! Пять или десять минут времени нужного человека могут окупить стоимость вашего билета.

Если вы живете в большом городе или рядом с ним, поищите конференцию там, чтобы не пришлось платить за проезд или проживание. В целом, если вы просите, чтобы компания заплатила за вас, лучшая стратегия — показать, как вы сможете оптимизировать работу благодаря знаниям, полученным на конференции.

14.2.1. Как справиться с социофобией

То, что инженеры-программисты и дата-сайентисты — социально не приспособленные интроверты, конечно же, стереотип, но, так или иначе, большинству людей в какой-то момент приходится бороться с социофобией. Даже самые уверенные в себе не заходят в комнату, полную незнакомцев, чувствуя себя при этом абсолютно комфортно. Что делать, если вы так сильно нервничаете и на конференции все время прячетесь в углу, уткнувшись в телефон?

К счастью, преимущество посещения таких мероприятий в том, что у вас есть общие темы для разговора! Задавать вопросы, в общем-то, хорошая стратегия: люди любят говорить о себе. Спросите, что привело человека сюда, как долго он программирует на X и бывал ли он на этих конференциях раньше. Помните, что неловко чувствуют себя многие, а не только вы. Если вы нервничаете, лучшее время, чтобы завязать разговор, — как раз за несколько минут до начала доклада. Заговорите, например, с соседом. Если из этого ничего не выйдет, то можете быть спокойны, ведь всего через несколько минут на сцену выйдет спикер!

Когда вы находитесь в помещении с толпой людей, идите к группе, стоящей в фигуре Пакмана: круг со свободным участком. Подойдите поближе к свободному участку и попытайтесь продвинуться ближе, чтобы освободить место для вновь прибывших. Не обязательно представляться сразу, как только подойдете: можно дождаться паузы. А можно не представляться вовсе, особенно если группа большая.

В главе 8 мы говорили о синдроме самозванца, и это еще одна сфера, где он может нанести удар. Вы можете попасть на слишком сложную презентацию, которая окажется вам не по зубам. Самое важное, что нужно помнить, — не чувствуйте себя самозванцем. Если к вам относятся так, как будто на вас не стоит тратить время, или кого-то шокирует, что вы не знаете термин, или если в отношении вас вообще проскакивают уничижительные комментарии, то это не ваша проблема.

Есть много конференций с дружественной атмосферой. Многие любят помогать другим людям и помнят, что значит быть новичком. Если вы оказались на конференции, которая вам не нравится, не дайте внутренним переживаниям убедить вас, что на такие мероприятия вы больше ни ногой.

В этой главе мы говорили в основном о стратегиях взаимодействия с другими людьми, но совершенно нормально выделять время на себя. Несколько дней общения с незнакомцами могут утомить. Распространенная ошибка — думать, что нужно обязательно делать что-то продуктивное каждую минуту конференции, будь то выступление или нетворкинг. Но вы совершенно не обязаны! Не переживайте из-за того, что проводите какое-то время в одиночестве вместо участия в разговоре; конференция принесет вам больше пользы, если у вас будет достаточно энергии.

14.3. Выступление с докладом

Выступление с докладом дает множество шансов для карьерного роста, а также возможность посещать больше конференций и слушать выступления большего числа специалистов. Одна из проблем, с которыми вы можете столкнуться, — это найти время, необходимое для участия. Но выступление с докладами это также отличная возможность рассказать о вашей компании (вы получаете бонус в виде премии за выступление, а ваше участие становится менее обременительным для работодателя). На первый взгляд может показаться, что для выступления обязательно нужно быть отраслевым экспертом, ярким оратором или очень общительным человеком, но все это вовсе не обязательно. На самом деле выступить с докладом — отличная стратегия для интроверта. После выступления к вам будут подходить люди, чтобы похвалить презентацию, задать дополнительные вопросы или просто познакомится. Сделать свое выступление темой для диалога — это улучшенная версия способа заговорить на общие темы конференции.

На эту тему можно написать целую книгу, и фактически в разделе материалов к главам 13–16 мы рекомендуем книгу о публичных выступлениях. Мы хотим подчеркнуть, что планка для хорошего выступления ниже, чем вы думаете. Вы не выступаете на TED и не делаете основной доклад на большой конференции. У участников такого уровня уже есть огромный опыт, и они наверняка нанимали специалистов по публичным выступлениям. Мы считаем, что если вы хотите хорошо выступить, то должны сосредоточиться на двух вещах: развлекать людей и мотивировать их. Если вас неинтересно слушать, то чему-то научить очень сложно. Кроме того, люди могут сделать слишком мало выводов из 20-, 30- или даже 60-минутного выступления. Но если вы сможете разжечь интерес, то принесете большую пользу.

14.3.1. Получение возможности

Как найти возможность выступить с докладом? Начните с конференций, на которых проводятся конкурсы заявок (calls for proposals, CFP). Вы можете подать краткое изложение доклада, которое называется *аннотацией*, на основании которой организаторы делают выбор. Некоторые конференции проводят слепой отбор, когда выбирают тезисы, ничего не зная о докладчиках.

Когда вы ищете конференции для выступления, выбирайте по тем же критериям, что и при поиске мероприятий, которые вы бы хотели посетить. Если конференция с участием 10 000 человек вызывает у вас приступ паники, вряд ли стоит подавать заявку туда. Более того, выступление в качестве докладчика — отличный способ снизить стоимость участия, так почему бы не воспользоваться этим и не отправить заявку туда, где вы действительно хотели бы послушать выступление других спикеров? Обязательно спрашивайте людей, что они рекомендуют; можно очень легко упустить отличные конференции небольшого формата.

Первое правило хорошей аннотации — учитывать требования конференции. Даже если вы напишете идеальную аннотацию из 500 слов, ее не примут, если по правилам она не должна составлять более 150. То же правило работает, когда вы отправляете доклад по инженерии данных на конференцию, посвященную статистике.

В хорошей аннотации первое предложение самое главное: оно привлекает внимание читателя, вызывая желание узнать больше. Затем вы должны объяснить решаемую задачу и сделать обзор того, что планируете рассказывать слушателям. Вот один из примеров, составленный для выступления Жаклин:

Глубокое обучение звучит сложно, но на самом деле это не так. Благодаря таким пакетам, как Keras, вы можете начать всего с нескольких строк кода. Как только вы освоите основные концепции, то сможете использовать глубокое обучение для создания юмористического контента на основе искусственного интеллекта! В этом выступлении я расскажу о глубоком обучении и покажу, как его можно использовать для создания модели, которая генерирует странные имена питомцев, такие как Шурпер, Тункин Пайк и Джек Одинс. Если вы знаете, как выполнить линейную регрессию, то поймете, как создавать увлекательные проекты глубокого обучения.

Когда у вас появляется некая абстрактная идея, вспомните, с чего вы начинали свой путь, что знали и умели три, шесть или двенадцать месяцев назад. Что вы знаете сейчас из того, что хотели бы знать тогда? Вы можете считать, что все и так знают эту тему, но то, что вы воспринимаете как базовые знания (например, как использовать git и GitHub или как выполнять парсинг веб-страниц), кому-то может показаться новым и весьма полезным. Можно также выбрать свою подобласть, если это будет интересно широкой аудитории. Может быть, вы сможете расска-

зать, как делать интерактивные карты, использовать пакет, специально предназначенный для быстрого анализа данных, или объясните, что такое обобщенная линейная модель. Вам не нужно быть экспертом в этой области; по сути, люди, которые только что чему-то научились, часто оказываются лучшими учителями. Те же, кто освоил это давным-давно, часто забывают, каково находиться в самом начале пути.

Еще один отличный способ начать выступать — местные митапы. Выясните, проводятся ли в вашем городе какие-либо мероприятия, где разные спикеры выступают с блиц-докладами (около пяти минут). К ним гораздо легче подготовиться, поскольку за один вечер выступает от пяти до десяти спикеров. Такие мероприятия отлично подходят тем, кто выступает впервые. Если ничего такого не планируется, но вы регулярно посещаете местные митапы, порекомендуйте организаторам провести подобный вечер с блиц-докладами!

Габриэла де Кьерос (Gabriela de Queiroz): создание R-Ladies

Когда я переехала в Сан-Франциско из Бразилии в 2012 году, количество открывшихся возможностей меня поразило. Я быстро нашла площадку для семинаров и в течение нескольких месяцев посещала их каждый день. Бесплатно есть и учиться — идеальное сочетание, особенно для студента, у которого мало денег. Но на большинстве встреч аудитория была однообразной. Я не видела никого, похожего на меня, и мне было некомфортно, поэтому в конечном итоге я оказывалась где-то в сторонке и мало общалась.

Через некоторое время я решила, что пора внести свой вклад, и начала сама вести семинары. Я была увлечена R, но не хотела создавать обычную группу по его изучению. Мне нужна была группа, в которой и я, и участники могли бы чувствовать себя комфортно и безопасно, не боясь осуждения и ощущая себя частью аудитории. Так родилась компания R-Ladies. В октябре 2012 года я провела первое мероприятие, знакомство с R (http://bit.ly/rladies-first), и на него пришло всего восемь человек. Я была немного разочарована, но все же радовалась тому, что создала это пространство, и мне хватило смелости, чтобы преподавать программирование на иностранном языке.

В течение четырех лет я была единственным человеком, стоящим за R-Ladies. Я занималась организацией, хостингом, обучением, рекламой, ведением веб-сайта, поиском мест и спонсоров. Я посещала конференции и другие мероприятия, где рассказывала о группе. Я была активна в соцсетях, старалась завязать как можно больше знакомств. К сожалению, большинство работодателей не спонсировали эту работу, поэтому R-Ladies не был для меня основным проектом: я работала над ним по ночам и в выходные.

Руководство R-Ladies дало мне возможность познакомиться с множеством людей, с некоторыми из которых я даже не мечтала встретиться в реальной жизни. И, поскольку мне приходилось преподавать на мероприятиях, мне стало гораздо легче выступать перед аудиторией. Людям, желающим создать собственное сообщество, я бы дала несколько советов:

- *Определите цель и сформулируйте миссию*. Какова цель этого сообщества? Чего вы пытаетесь достичь? Зачем вы его создаете? В чем миссия сообщества? Кто в него вступит? Ответы на эти вопросы помогут вашим будущим участникам понять, почему это сообщество достойно внимания и зачем к нему стоит присоединиться. Они также помогают оценить, стоит ли вам сосредоточиться на определенной подгруппе, как сделали R-Ladies с женщинами и гендерными меньшинствами, или же вы хотите охватить всех, кто интересуется этой темой.
- *Создайте каналы в соцсетях, веб-сайт и электронную почту*. Создайте учетную запись в Twitter, страницу в Facebook, группу в LinkedIn, профиль в Instagram и канал в любой другой социальной сети с большой пользовательской базой. У вас должны быть веб-сайт и электронная почта, чтобы люди могли легко связаться с вами и узнать больше о группе.
- *Создайте логотип*. Наличие логотипа повышает узнаваемость вашего бренда и, следовательно, вашего сообщества. У некоторых людей визуальная память развита лучше, и они запомнят ваш логотип. Например, вы можете сделать брендированные наклейки для ноутбука — это отличный способ выразить себя, свои убеждения и показать свою причастность к сообществу. Это большой успех!
- *Продумайте формат*. Будут ли это преимущественно лекции или семинары? Станете ли вы онлайн-сообществом с мероприятиями в прямом эфире или будете проводить очные встречи за чашкой кофе? Если это будет техническое сообщество, цель которого — способствовать развитию аудитории, формат семинара будет оптимальным. Активное обучение — лучший способ чему-то научиться.
- *Используйте платформу* (например, meetup.com или eventbrite.com). Вы должны облегчить людям поиск и регистрацию на ваши мероприятия. Специализированные веб-сайты вроде www.meetup.com или www.eventbrite.com обеспечат естественный трафик, когда люди будут искать интересную им тему, и поможет оценивать ожидаемое количество посетителей.

Создание сообщества требует времени и усилий. Возможно, вам придется работать в свободное время и в выходные, поэтому убедитесь, что вас это действительно вдохновляет и вы верите в миссию сообщества. Несмотря на все сложности, оно того стоит. Услышать истории успеха, увидеть, как ваше сообщество повлияло на людей по всему миру, особенно в местах, недостаточно обеспеченных услугами, — это очень ценный источник радости и воодушевления. Вы чувствуете, что попытались сделать мир лучше. Удачи вам!

Наконец, вы можете сделать так, чтобы люди выходили на вас через ваш блог. Что касается конференций, куда приглашают спикеров, если один из организаторов прочитает пост в блоге, который идеально соответствует теме мероприятия, он может связаться с вами и уточнить насчет вашей возможности сделать доклад. Если вы не можете предоставить ссылки на свои предыдущие доклады, посты в блогах — отличный способ показать организаторам конференции, что вы умеете эффективно передавать информацию.

Получить шанс на первое выступление так же непросто, как и первую работу в Data Science. Далее ваш опыт будет расти как снежный ком, особенно если доклад выложат в записи. Здесь плюс в том, что другие смогут увидеть ваше выступление и связаться с вами при необходимости, а также это здорово потому, что в некоторых случаях организаторы просят дать ссылки на подобные видео.

14.3.2. Подготовка

Перед выступлением вы должны потратить много времени на подготовку доклада. Если вы никогда раньше не выступали на публике, то можете легко недооценить количество часов, которое уходит на это дело. Да, можно было бы написать речь в последнюю минуту, создав презентацию, где на каждом слайде было бы по пять пунктов, отражающих главные идеи. Но это неуважительно по отношению к аудитории и не демонстрирует вас с лучшей стороны. Это плохой путь к успешному ораторству.

Вы должны потренироваться выступать перед другими, а не просто прочитать слайды про себя. Найдите кого-нибудь, чьей критике вы доверяете, и выступите с докладом перед этим человеком. Если у вас не очень технический доклад, то экспертиза слушателя не имеет значения. Вам могут дать общие советы о том, как улучшить выступление, например посоветовать сократить количество слов-паразитов или жестикулировать менее активно.

Обычно для выступления отводится определенное время, которое зависит от того, будут ли в конце задаваться вопросы. Чтобы рассчитать, сколько времени вам понадобится, берите запас в пять минут на вопросы и работайте в обратном направлении; засеките, во сколько укладываетесь. Учтите, что во время публичного выступления может случиться так, что вы будете говорить быстрее, чем следовало, и закончите доклад раньше. На этот случай добавьте несколько запасных слайдов в конце презентации. Не страшно, если доклад затянулся на несколько минут: тогда между вашим и следующим выступлением будет перерыв подлиннее. В самом худшем случае доклад окажется слишком длинным, и вы либо оборвете его, не закончив, либо слишком сильно задержите следующего выступление.

Над любым докладом требуется много работать, поэтому мы настоятельно рекомендуем не писать новый каждый раз. Очень маловероятно, что аудитория будет пересекаться, особенно если лекции проходят в разных городах или проводятся на «многодорожечной» конференции (где участники могут выбрать параллельное выступление). Мысль, что запись вашего выступления посмотрели все, греет душу, но в большинстве случаев такого не бывает.

Пригласите на свое выступление группу поддержки. Это не обязательно должны быть друзья и коллеги из Data Science: пригласите членов семьи, своего партнера и вон того милого соседа. Если мероприятие платное, узнайте, можно ли получить пропуск на этих людей. Дедушка Эмили побывал на нескольких ее

выступлениях бесплатно (к большому удовольствию других присутствующих). Приятно осознавать, что хотя бы часть аудитории поддерживает вас.

14.4. Вклад в открытый исходный код

Для тех, кто любит быть частью сообщества, но не в восторге от идеи нахождения в одном помещении с кучей людей, открытый исходный код может стать отличным решением. Участие в его разработке позволяет обмениваться идеями и развивает чувство общности среди специалистов с одинаковыми интересами. Это занятие может быть очень интересным, поскольку связано с множеством новых направлений, о которых вы, возможно, не задумывались. Точно так же вы можете использовать чью-то работу, чтобы создать совершенно новый проект.

R и Python успешно развиваются, потому что волонтеры постоянно расширяют и совершенствуют их. В следующих разделах мы обсудим, как стать одним из этих людей; а еще можно внести финансовый вклад в организации, которые спонсируют некоторые из основных разработок. Хотя R и Python можно использовать бесплатно, с их поддержкой и разработкой дело обстоит иначе. R Foundation, Python Software Foundation и NumFOCUS — это три организации (последние две зарегистрированы в США), которым вы можете отправлять деньги для поддержки дальнейшего развития этих языков.

14.4.1. Участие в работе других людей

Начиная проект с открытым исходным кодом, может показаться, что вы заглянули в чей-то шкаф. Это чужая территория, и вы чувствуете себя злоумышленником, но open source был создан именно для этой цели, так что нужно избавиться от подобного чувства. Представьте, что проекты с открытым исходным кодом похожи на огромный званый ужин. Наверное, пока не стоит брать на себя ответственность за приготовление основного блюда, но есть еще много другой работы, в которой вы можете поучаствовать, к примеру помочь накрыть стол, убедиться, что у всех есть вода, или убрать тарелки. Если вы проявляете уважение и энтузиазм, большинство создателей и разработчиков будут рады вашей помощи.

Работа с документацией — отличный вариант для старта. Посмотрите, насколько укомплектована документация для понравившегося вам пакета. Вы можете заметить, что что-то указано не полностью, написано непонятно или вводит в заблуждение. Даже исправив опечатку, стоит сделать запрос на включение внесенных изменений на GitHub. Создатели пакетов и библиотек любят, когда им помогают. Это экономит их время, а вы, как человек, недавно научившийся использовать эти инструменты, лучше поймете, что будет мотивировать новых пользователей и чему они научатся.

Если вы хотите внести свой вклад в код, не спешите сразу переписывать что-то или отправлять новую функцию. Если проект большой, у него могут быть управляющие или определенные правила. Если это не так, понаблюдайте за репозиторием, чтобы понять, как он работает. Наблюдение также покажет, насколько активно поддерживается проект. Если вы решили, что хотите начать писать код, для начала поделитесь тем, что хотите добавить или изменить. Таким образом, вы сможете получить обратную связь еще до того, как проделаете большую работу.

Решама Шейх (Reshama Shaikh): спринты

Работа над открытым исходным кодом может показаться малопонятной и внушающей страх. Соревнования по open source, иногда называемые спринтами, представляют собой организованные мероприятия, которые создают благоприятную атмосферу для новичков. Обычно они длятся один или два дня, когда участники работают над открытыми задачами, отправленными в репозиторий GitHub библиотеки Python или R. Эти задачи могут быть связаны с документацией, исправлениями багов, тестами, запросами функций и так далее.

Участие в спринтах по открытому исходному коду дает много преимуществ:

- Большинство участников — волонтеры, поэтому общественная работа востребована и приветствуется.
- Это активное практическое мероприятие, которое развивает навыки инженерии и программирования.
- Работа с открытым исходным кодом — это отличная возможность обучения, которая развивает навыки в Data Science и помогает создать портфолио.
- Спринты дают ценную возможность наладить контакт с другими дата-сайентистами и прочими опытными участниками.

Хорошо организованный хакатон позволит эффективно использовать время людей. Подготовка гарантирует, что новички вынесут для себя что-нибудь полезное и добьются хороших результатов. Поищите доступный центральный репозиторий ресурсов и подготовительных работ, который включает сопутствующую документацию, инструкции по установке R или Python, инструменты для предварительной подписки на мероприятия (например, учетную запись GitHub или платформу обмена сообщениями) и список открытых задач, специально подготовленных для участников спринта. Имейте в виду, что организаторы тоже волонтеры; если вы обнаружите, что чего-то не хватает, предложите помощь. Организация спринтов по open source — это тоже важный вклад.

Целью спринтов является отправка запросов «pull-requests» (PR), которые решают открытые вопросы. Отправка PR — это двусторонний процесс, и обычно для их слияния требуется несколько недель. Выделите некоторое время после спринта (обычно от 5 до 10 часов), чтобы проследить за работой и увидеть PR вплоть до статуса слияния, который представлен в репозитории GitHub красивым фиолетовым значком.

Если вы интересуетесь самостоятельной организацией спринта, ознакомьтесь с подробным руководством в моем блоге по ссылке https://reshamas.github.io/how-to-organize-a-scikit-learn-sprint.

Участие в проектах с открытым исходным кодом — один из лучших способов развить технические навыки, особенно если от вас не требуется взаимодействовать с большой группой людей. Возможно, в своих рабочих репозиториях GitHub вы не используете ветки, сообщения о завершении транзакции или теги — это нормально. Но когда вы приходите в проект с сотнями задач и десятками исполнителей, эта дополнительная работа обретает больше смысла. Подобные методы действительно накладывают некоторые ограничения независимо от того, работаете ли вы над руководством по стилю или же персонал сопровождения не добавляет созданную вами функцию из-за ее недостаточной производительности. В конечном итоге кто-то будет принимать окончательные решения, пока вы не создадите собственный проект. Хотя это может быть непросто, вы узнаете множество передовых методов, которые можно применить в работе.

14.4.2. Создание собственного пакета или библиотеки

Когда вы понимаете, что копируете функции разных проектов или отправляете их коллегам, возможно, пришло время создать пакет или библиотеку. Пакет позволяет хранить функции в одном месте, легко делиться ими и применять современные методы, такие как тестирование кода. У многих компаний есть внутренние пакеты с функциями, позволяющими выбрать корпоративную расцветку графиков, получить доступ к данным или решить общие задачи. Если вы думаете, что у других людей могут возникнуть аналогичные вопросы, поделитесь своим пакетом на GitHub, чтобы они могли его загрузить и использовать.

> **ТОКСИЧНОСТЬ В OPEN SOURCE** Сообщества с открытым исходным кодом могут быть токсичными. У многих людей был негативный опыт, когда их дискриминировали, домогались, унижали или просто заставляли чувствовать себя неугодными из-за расы, пола, этнической принадлежности или сексуальной ориентации. К счастью, многие сообщества осознают этот факт и активно работают над созданием атмосферы с учетом индивидуальных особенностей. Гвидо ван Россум (Guido van Rossum), автор Python, взял на себя обязательство обучать только женщин и недостаточно представленные меньшинства (http://mng.bz/9wPo). Некоторые создатели проектов ставят тэги «рассчитанный на начинающих» или «новичок», чтобы вдохновить тех, кто не знаком с открытым исходным кодом, внести свой вклад. Принимая во внимание, что на первом месте всегда должно стоять психическое и эмоциональное здоровье, многие люди, в том числе из недостаточно представленных социальных групп, имели только положительный опыт работы с open source; но и негативный тоже бывает.

Прежде чем предлагать кому-то использовать ваш продукт, убедитесь, что с кодом все порядке. То, что у вас что-то хорошо работает в одной задаче, не

означает, что код выдержит нагрузку массового потребления. Если вы слепили код из того, что было, но не знаете, как он работает, не предлагайте его другим. Чтобы ваш пакет можно было использовать, может потребоваться более продвинутое программирование, поскольку вы дорабатываете или адаптируете его для обобщенных случаев. Убедитесь, что проделанную вами основную работу проверил специалист, которому вы доверяете. Пользователи не будут заглядывать под капот, поэтому, если вы скажете им, что это Ferrari, они сильно расстроятся, когда каждый второй раз это будет оказываться машина для гольфа.

Когда вы протестировали и проверили код, нужно еще постараться, чтобы о нем узнали. Вы можете продвигать его в социальных сетях или в блоге, но даже в этих случаях процесс может идти медленно. Не ждите, что сразу станете звездой; пусть поначалу у вас будет меньше пользователей — это лучше, чем если выстрелить сразу, а затем понять, что где-то в коде затесалась ошибка. Пользователям требуется время, чтобы принять что-то новое, если это вообще случится, но даже просто попытка поделиться хорошей работой — уже благое дело. Вознаграждение за успех одновременно и проклятие: если ваш проект начнет пользоваться спросом, будет сложнее прекратить его разработку. Вы будете получать отчеты о багах и запросы на новые функции, так что придется серьезно думать над тем, собираетесь ли вы вносить изменения, которые нарушат отчеты, где использовалась старая версия функции.

14.5. Распознавание и предотвращение выгорания

Мы не являемся экспертами в области здравоохранения, поэтому берем за основу определение выгорания, данное Всемирной организацией здравоохранения: «Синдром, признаваемый результатом хронического стресса на рабочем месте, который не был успешно преодолен». Он характеризуется тремя признаками: «ощущение мотивационного или физического истощения», «нарастающее психическое дистанцирование от профессиональных обязанностей или чувство негативизма или цинизма к профессиональным обязанностям» и «снижение работоспособности» (https://www.who.int/mental_health/evidence/burn-out/ru). На данный момент мы сосредоточимся на стрессе, который возникает не на основной работе, а из-за сторонних проектов.

Мы писали эту книгу вне основной занятости (а Жаклин еще и отдельно от воспитания малыша). Мы, конечно, иногда завидуем коллегам, которые вечером идут домой и не делают ничего, связанного с Data Science. В нашем случае это помогает вернуться к вопросу о том, почему мы решили взять на себя эту дополнительную работу, и понять, работаем ли мы по-прежнему ради достижения наших целей. Мы писали эту книгу совершенно не ради денег. (Консультации были бы намного прибыльнее.) Скорее мы хотели помочь начинающим специалистам по данным,

в этом была наша мотивация. Было особенно приятно видеть положительный результат наших стараний, по мере того как мы выпускали главы.

Если вы чувствуете, что выгораете, сначала спросите себя, есть ли возможность сократить количество работы. Помните, что после создания чего-нибудь не обязательно поддерживать активность постоянно. Если вы ведете блог, можете периодически писать новые статьи, но этого не стоит делать так часто, как в самом начале. Ваши читатели с большей вероятностью скажут: «Ого, эти шесть постов действительно мне помогли», а не: «О, теперь она пишет что-то только раз в полгода».

В современной культуре подработок и хвалебных речей в честь дополнительных часов на рабочем месте может казаться, что любое время на себя потрачено впустую. Это очень опасно! Нам всем нужна перезагрузка. Поход в тренажерный зал и тусовки с друзьями отлично подходят для этого, но можно и просто посмотреть телевизор или вообще бездельничать. Продолжайте уделять время хобби, совершенно далекие от Data Science или зарабатывания денег, чтобы не казалось, будто вся ваша жизнь вращается вокруг работы.

Попытки угнаться за другими могут выматывать не меньше. Часто говорят о том, что социальные сети — это нарезка лучших моментов из жизни других людей, поэтому не стоит сравнивать их со своей. Точно так же то, что кто-то активно занимается созданием пакетов или публикацией статей, не означает, что вам нужно идти в ногу с этим человеком. Для некоторых такая деятельность составляет часть основной работы, если не всю! Лучшее, что вы можете сделать для своей карьеры, — это убедиться в сохранении баланса между работой и жизнью, чтобы вас хватило на дольше.

14.6. Интервью с Рене Teate, директором отдела Data Science в HelioCampus

Рене Teate (Renee Teate) известна в Twitter под ником Data Science Renee (@becomingdatasci), где у нее более чем 50 000 подписчиков. Она также является создателем подкаста, блога (www.becomingadatascientist.com) и онлайн-справочника по Data Science (www.datasciguide.com). Рене регулярно выступает и организует конференции.

Каковы основные преимущества социальных сетей?

Twitter очень сильно мне пригодился. На самом деле всех гостей своего подкаста я встретила именно на этой платформе. Я нашла людей, которые, как мне казалось, писали классные твиты, и решила, что если они могут писать об интересных вещах, то говорить смогут ничуть не хуже. Я составила список и отправила сразу несколько сообщений в директ в надежде, что половина из получателей заинтересуется моим предложением. В итоге согласились все!

Меня постоянно просят выступить на конференциях и митапах через Twitter. Когда я публикую контент, то знаю, что у него есть определенная аудитория, которую он заинтересует. Я познакомилась со многими крутыми людьми. Помимо нетворкинга я также использую социальные сети для обучения. У меня есть пост в блоге, где я рассказываю, как использую Twitter для изучения нового: в нем в основном говорится о том, как пополнить свой словарь профессиональных жаргонизмов. Если вы подпишитесь на специалистов определенной отрасли и начнете читать их статьи, то вскоре усвоите всю эту терминологию. Если кто-то рассказывал о чем-то новом для меня, я просто искала информацию сама и выясняла, о чем шла речь. Часто к статье прилагалась ссылка на туториал или публикацию по теме, так что это действительно помогало в обучении.

Что бы вы сказали тем, кто говорит, что у него нет времени на взаимодействие с сообществом?

Я действительно их понимаю, особенно тех, у кого есть другие обязанности, помимо работы: например, уход за ребенком или другим родственником. Когда я училась в магистратуре и работала полный день, то больше ничем не занималась. В таких случаях я бы посоветовала найти онлайн-сообщество, с которым вы можете взаимодействовать асинхронно. Когда у вас есть немного времени, например пока кого-то ждете, вы можете увидеть и ответить на несколько твитов или добавить в закладки интересные статьи, чтобы прочитать их позже. Даже если вы можете посещать только одно мероприятие в год, выберите конференцию, которая связана с Data Science или с конкретной отраслью, и постарайтесь втиснуть ее в свой плотный график. Можно поддерживать связь в LinkedIn или других социальных сетях с новыми знакомыми с таких мероприятий, и в конечном итоге эти люди могут стать той группой, с которой получится делиться ресурсами.

Есть ли смысл в создании небольшого количества контента?

Конечно, да. Я думаю, что даже если вы сделаете лишь одну публикацию в блоге, она поможет закрепить знания темы, потому что вы учитесь сами, когда помогаете другим людям что-то освоить. Я много лет ссылалась на некоторые из своих старых постов. Когда я веду блог, то стараюсь сделать его практичнее и полезнее, а также слежу, чтобы к нему можно было вернуться несколько раз и чтобы при этом он не терял своей актуальности. Что касается моего подкаста, то я записала только две серии за последние полтора года. Я была занята, к тому же начала работать в HelioCampus и на время отложила этот проект в сторону; мне было действительно трудно возобновить его. Записи все еще выложены в блоге, и в моих планах есть еще несколько серий. Сейчас я перестала чувствовать себя виноватой из-за длительного перерыва. Я понимаю, что вышедшие эпизоды по-прежнему кому-то нужны, и я всегда могу вернуться к ним позже.

Вы волновались, когда впервые публиковали пост в блоге или выступали с докладом?

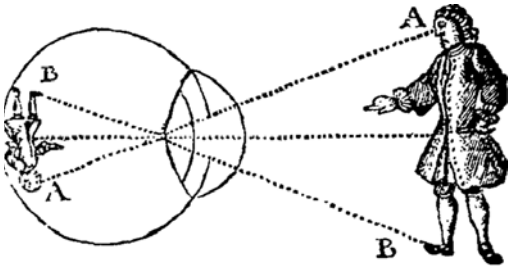
Да, мне очень хотелось опубликовать что-то, потому что когда люди будут искать ваше имя, они наткнутся на вашу статью или работу, с чем и будут вас ассоциировать. Конечно, это немного волнующе. Но я поняла одну вещь: прошлые публикации, которые я ценила, не всегда были написаны хорошо и не были технически передовыми. Я читала блоги, в которых то, что я пыталась понять раньше, описывалось по-другому, и внезапно пазл в голове складывался. Всегда найдется кто-то, кому пригодятся ваши материалы.

Я также научилась не обращать внимания на скептиков. Они будут всегда. Я видела негативные отзывы о специалистах, которые годами занимались Data Science. Есть много способов подойти к анализу, и порой какой-то из них лучше другого, но это не значит, что вы выбрали плохой путь. Иногда просто нужно перестать слушать недоброжелателей.

Итоги

- Мы рекомендуем четыре способа стать частью сообщества дата-сайентистов: создание своего Data Science блога и портфолио, посещение конференций, выступление с докладами и участие в разработке открытого кода.
- Помните, что вам не обязательно заниматься какой-либо общественной деятельностью, чтобы построить успешную карьеру; выбирайте то, что вам подходит, и не стремитесь идти в ногу с кем-то другим.

15



Уходим красиво

В этой главе

- Как принимать решение об увольнении.
- Разница между нынешним и первым поиском работы в Data Science.
- Предупреждение о предстоящем увольнении и управление переходом.

Прошли те дни, когда мы проводили по 40 лет в одной компании, а при выходе на пенсию выдавались золотые часы и пособие. В большинстве областей сейчас принято менять место работы минимум несколько раз за всю карьеру, а в сфере технологий специалисты и вовсе могут делать это каждые пару лет. Есть много веских причин для ухода с работы: может захотеться зарплаты повыше, других обязанностей, обучения или просто чего-то нового. Мысли о новой работе — это первый шаг, но прежде, чем перейти к действиям, нужно преодолеть дополнительные психологические барьеры.

Сомнения в том, стоит ли менять старое место, где все так хорошо знакомо, будут всегда. Сколько бы исследований вы ни провели или сколько бы вопросов ни задали на интервью, невозможно узнать заранее, как все будет на самом деле. У вас появится представление о важных составляющих — зарплате, размере компании и структуре DS-команды, но вы не узнаете, как будете чувствовать себя каждый день, пока не проживете его. Более того, вполне вероятно, что нынешняя

работа не так уж и ужасна. (Если это так, мы рекомендуем вернуться к главе 9, где рассказываем, что делать, если вы попали на ужасную работу или в токсичную среду.) Вам могут нравиться некоторые коллеги, вы знаете, у кого можно попросить помощи, и в целом чувствуете себя комфортно, занимаясь данными. Вам может хотеться чего-то большего. Но какова гарантия того, что новая работа, которая, по вашему мнению, будет лучше, на самом деле не будет еще хуже? Стоит ли рисковать и тратить время на поиск нового места?

Эти мучительные сомнения могут замедлить поиск работы даже после решения уволиться. Наверняка вы не знаете, что делать. С чего начать поиск во второй раз? Если вдруг вы получите оффер, который захотите принять, то как сообщить об этом своему руководителю? Стоит ли лично говорить об этом каждому коллеге, с которым вы работали над проектом? Если вам сделают встречный оффер, согласиться на него? Что делать в последние несколько недель, после того как вы сообщили об увольнении? Но поиск второй (третьей и даже четвертой) работы не должен вас смущать, ведь желание найти место, где трава зеленее, вполне может измениться.

Мы собрали много вопросов, которые задавали себе, когда задумывались о поиске новой работы. Огромная неуверенность может парализовать всех, кроме самых стойких соискателей, но не бойтесь: мы здесь для того, чтобы превратить вас в одного из них.

В этой главе мы изящно делим процесс на три части: решение уйти, начало поиска нового места и предупреждение о предстоящем увольнении. Некоторые из этих рекомендаций применимы к любой работе, но мы также обсудим некоторые моменты, которые характерны в большей степени для Data Science. Смена работы в этой сфере — обычное дело и полезный опыт; многие специалисты меняют работу каждые пару лет, что позволяет им пробовать новое, значительно повышать зарплату и получать другие преимущества. Эта глава поможет вам сделать такой переход максимально легким и комфортным.

15.1. Решение уволиться

К сожалению, в большинстве случаев вы не будете знать со стопроцентной уверенностью, когда лучше всего уйти. Не существует магического шара, который подскажет вам, что делать, равно как нет и вопросов, на которые вы можете ответить, чтобы принять окончательное решение. В главе 8 мы говорили о том, как выбирать между двумя хорошими офферами; в этом случае работают те же принципы. В конце концов, вы можете сделать все возможное, используя имеющуюся у вас информацию. И помните, что очень немногие решения абсолютно необратимы. Уволиться можно всегда; вы не подписываете 100-летний контракт.

15.1.1. Оценка прогресса в знаниях

Что должно послужить сигналом к поиску новой работы? Наш самый главный совет — убедитесь, что вы всегда чему-то учитесь. К сожалению, в этом плане может происходить застой, если вы надолго задерживаетесь на одной и той же должности. В первые несколько месяцев на новом месте вы словно пьете воду из пожарного шланга. Практически невозможно *не* узнать чего-нибудь полезного: вы изучаете данные компании, получаете новые технические навыки и работаете с бизнес-стейкхолдерами. Но если вы продолжите делать то же самое через год или два, то можете выйти на плато.

Освоившись с ежедневными обязанностями, подумайте, как улучшить свои нетехнические навыки. Посмотрите, можете ли вы взять на себя ответственность за команду (или хотя бы за стажера) и поработать над управленческими способностями. Несмотря на то что работа, на которую вас нанимали, может в какой-то степени ограничивать деятельность, по мере получения опыта у вас появляется больше времени на то, чтобы расширить свое портфолио. Возможно, вы могли бы сотрудничать с командой DS-инженеров, чтобы научиться строить часть конвейера самостоятельно, а не полагаться исключительно на них. Однако проявление инициативы для самопродвижения подходит не всегда; иногда, чтобы люди решали новые задачи, им требуется внешняя мотивация от компании. Если вы попали в колею и не можете из нее выбраться, это знак, что, возможно, пришло время сменить обстановку.

В Data Science замечательно то, что всегда есть чему учиться, но этот факт также усложняет работу. Если вы не растете, перейти на ступень выше будет сложнее. Предполагается, что обширность и глубина знаний синьора будут заметно отличаться от навыков джуна. На протяжении всей книги мы подчеркивали, что не нужно — да и невозможно — знать все, но по мере накопления опыта знаний все же должно прибавиться.

15.1.2. Заручитесь поддержкой руководителя

Прежде чем сжигать мосты, убедитесь, что вы сделали все возможное, чтобы сообщить своему руководителю, что вы хотите изменить. То, что вам может показаться неразрешимой проблемой, на самом деле может иметь простое решение. Возможно, вы погрязли в легких рутинных задачах, которые, к сожалению, нельзя автоматизировать. Руководитель может предложить вам взять стажера для такой работы. В таком случае стажер получает опыт работы, а вы немного разгружаете себя и получаете опыт наставничества. Или может быть так, что команда по работе с данными занимается в основном аналитикой, а вы хотите осваивать производственное машинное обучение. Ваш руководитель может организовать для вас

«курс молодого бойца» с командой инженеров на несколько месяцев; вы изучите некоторые основы инженерии, поделившись своими аналитическими знаниями.

Еще один вопрос, который стоит задать себе, — насколько ваши цели совпадают с целями руководителя. Филип Гуо (Philip Guo), доцент кафедры когнитивных наук Калифорнийского университета в Сан-Диего, опубликовал статью *Whose critical path are you on?* (http://www .pgbovine.net/critical-path.htm), где говорит, как важно понимать критический путь своего руководителя (или наставника) и того, соответствует ли он вашему критическому пути. *Критический путь* означает «путь работы, который имеет решающее значение для карьерного роста или достижения цели в данный момент». Дело в том, что ваш успех связан с успехом руководителя. Обычно менеджер ограничен во времени и силах, и если ваши критические пути совпадают, то он с большей вероятностью сосредоточится на вас.

Увольнение «в никуда»

Возможно, вам нужен значительный перерыв. Большинство новых работодателей захотят, чтобы вы вышли как можно скорее; хотя обычно дается неделя или две недели между уходом с текущей работы и началом новой (особенно если у вас уже запланирован отпуск), вряд ли вам дадут больше. Если вы мечтали о трехмесячном туре по Азии, то, скорее всего, придется уволиться без каких-либо других планов.

Оставаться без работы рискованно, так как есть вероятность, что ваших сбережений окажется недостаточно, а вы остаетесь без дохода на неопределенное время. Готовы ли (или можете ли) вы занять денег у членов семьи или быть на содержании у партнера? Кроме того, проще найти работу, будучи уже трудоустроенным. Во-первых, так происходит из-за несправедливых предубеждений в отношении безработных, бытующих среди рекрутеров. Во-вторых, ваша позиция на переговорах будет слабее: любой оффер нового работодателя будет выгоднее текущего, поэтому вам будет сложнее требовать более высокую зарплату. Кроме того, если вы взяли паузу на несколько месяцев, то некоторые навыки могли не сохраниться и на технической части интервью вы поймете, что мозг немного заржавел.

Если вы действительно хотите сделать перерыв, то в этой ситуации очень пригодится нетворкинг: люди, которые знакомы с вашей работой и могут организовать встречу с рекрутером, будут очень кстати. Освежите свои технические знания перед собеседованием. В целом, если вы не оказались в токсичной рабочей среде, мы рекомендуем не увольняться в никуда без запасного плана.

Чтобы понять, насколько хорошо вы совпадаете со своим руководителем, необходимо знать собственные карьерные цели. Мы не говорим о десятилетнем или даже пятилетнем плане; в такой новой и быстро развивающейся области невозможно знать, какие возможности будут доступны в таком далеком будущем. Но как вы хотите провести следующие несколько лет? Мы надеемся, что вы много

размышляли над этим вопросом во время первого поиска работы, но, быть может, все изменилось. Возможно, вы хотели стать частью большой DS-команды, но через несколько лет поняли, что вам больше нравится работать над разными типами проектов, а не изолированно. Или вы хотите посвятить время семье и вам нужна работа с 9 до 5, а не стартап, требующий гораздо больше времени.

Итак, некоторые ключевые факторы, которые следует учитывать при поиске новой работы:

- Учитесь ли вы на своей нынешней должности?
- Пытались ли вы улучшить рутинные дела, обсуждали ли это с руководителем?
- Учитывает ли руководитель ваши потребности и желание продвигаться по карьерной лестнице?
- Задумывались ли вы о том, чего хотите и не хотите на следующем месте?

15.2. В чем разница между первым и последующими поисками работы

Многие принципы очередного поиска работы остаются теми же, что и для первого. Но вместе с опытом в Data Science у вас появляются некоторые существенные преимущества:

- Вы будете интересны большему количеству рекрутеров (любых). Чтобы они заметили ваше резюме, сделайте его доступным для просмотра в профиле LinkedIn (не беспокойтесь, LinkedIn принимает соответствующие меры, чтобы текущие работодатели этого не видели).
- Теперь вы лучше понимаете, какие аспекты работы вам нравятся, а какие нет. Менять специализацию вам еще рано: если вы много работали над инженерией данных, но вам не понравилось, поищите более крупную компанию, в которой есть инженеры по обработке данных для выполнения таких задач.
- Будет легче добраться до первого этапа отбора. Многие работодатели обращают внимание на то, занимал ли кандидат такую же или похожую должность.
- В идеале ваша сеть контактов станет больше. (Если нет, вернитесь к главе 14.)
- Если вы все еще работаете, но не хотите указывать в LinkedIn или Twitter, что находитесь в поиске, можете прозондировать почву и сообщить эту новость нескольким доверенным людям. Есть вероятность, что они порекомендуют вам вакансию в своих компаниях или свяжут вас с тем, кто ищет кандидатов.

Не бойтесь откликаться на вакансии, даже если вы в целом довольны своим нынешним положением. Есть множество причин, по которым вы можете отговаривать себя от перемен. Может быть, вы думаете, что у вас нет навыков, которыми,

по вашему мнению, обладают все остальные («Что, если я не пройду техническую часть и не справлюсь с кейсом, несмотря на свой опыт?»). Это просто синдром самозванца (и если вы что-то вынесли из этой книги, то пусть это будет умение бороться с назойливым внутренним голосом, говорящим, что вы не так хороши, как все остальные). Если вы не прошли отбор по техническим параметрам, это не значит, что вы неудачник или самозванец. Множество плохих вопросов на интервью не позволяют объективно оценить способности. Кроме того, Data Science настолько обширна, что, возможно, вас просто спрашивали о том, с чем вы раньше не работали.

А может, вас волнует, как смена работы повлияет на вашу социальную и семейную жизнь. Но какими бы ни были ваши опасения, вы окажете себе медвежью услугу, если не будете открыты для поиска новых возможностей, которые могут стать трамплином в карьере.

15.2.1. Определитесь, чего хотите

Первое, что нужно сделать при поиске новой работы, — составить список того, что вам нравится на текущем месте. Билл Бернетт (Bill Burnett) и Дэйв Эванс (Dave Evans) в книге *Designing Your Life: How to Build a Well-Lived, Joyful Life* (Knopf, 2016) предлагают в течение недели анализировать свои задачи, оценивая то, насколько они вам понравились в сравнении с вашими ожиданиями. Вы ненавидите ежедневные двухчасовые совещания или они вам нравятся, потому что структурируют рабочий день? Если вы работаете в распределенной DS-команде, хотите ли вы отчитываться ее менеджеру? Вы можете использовать этот список как образец. Найдите компанию с теми же ценностями или структурой, которые вы ищете. Не стоит обращаться в фирму, где есть те же проблемы, которые вы наблюдаете сейчас.

При поиске работы вы можете столкнуться с проблемой названия должности. Мы говорили об этом в главе 5. Дата-сайентисты могут называться по-разному: аналитик данных, инженер-исследователь, инженер по машинному обучению, аналитик продукции. Аналитик данных — наиболее распространенное название и может рассматриваться как должность младшего специалиста. Если вы дата-сайентист, готовы ли вы принять должность старшего аналитика данных? Если вы аналитик данных, стоит ли вам сосредоточить усилия на повышении до должности специалиста по данным в новой роли?

Обучение по-прежнему является самым важным фактором при поиске. Чем вы будете заниматься на новой должности? Думайте о перспективе на пять лет, а не на два года. Что обеспечит вам успех в долгосрочной перспективе? Можно ли прийти в компанию как старший аналитик данных, а затем получить должность дата-сайентиста? Позволит ли вам опыт в небольшой технологической компании научиться работать с веб-данными и перейти в крупную технологическую корпорацию?

Рассматривая варианты, вы должны отстоять свою рыночную стоимость. Справедливо или нет, но дата-сайентист по-прежнему считается более престижной должностью, чем аналитик данных, а зарплата последнего на уровне синьора может быть ниже, чем у первого. Учитывайте эти факторы.

15.2.2. Интервью

На интервью вас спросят: «Почему вы решили уволиться?» Если на нынешнее место работы вы пришли после университета или курсов, то такой вопрос вам вообще не задавали.

Будет отлично, если вы скажете, что хотите решать более сложные задачи. Еще одна хорошая стратегия — представить, что вопрос звучит так: «Почему вы хотите работать у нас?» Если сделать это, то ответ будет звучать в положительном ключе («Я слышал много хорошего о вашей команде машинного обучения и очень хочу стать ее частью»), а не в отрицательном («Мой бывший начальник требовал представлять результаты эксперимента в виде круговых диаграмм»). Если вы отвечаете более конкретно, убедитесь, что новому работодателю ответ понравится. Не стоит говорить: «Я ищу команду, где смогу работать со старшими дата-сайентистами», если у компании, проводящей интервью, таковых нет! Любой ценой избегайте говорить о текущем работодателе плохо; некоторые рекрутеры такого поведения не признают независимо от того, сколько правды в ваших словах.

Тот факт, что вы уходите из компании, не означает, что вы не должны гордиться проделанной там работой. Обязательно говорите о проектах, в которых участвовали, или о приобретенных навыках. Скорее всего, вы будете ограничены определенными соглашениями о конфиденциальности, поэтому не сможете наглядно продемонстрировать свой код или рассказать о параметрах созданного алгоритма рекомендаций, но вам следует обсудить свой вклад в общих чертах. Прекрасный обтекаемый ответ: «Я создал чат-бота на Python, который генерировал ответы на частые вопросы клиентов, что уменьшило среднее время, необходимое представителю службы поддержки для общения, на пять минут и увеличило удовлетворенность клиентов на 20 %». Если вы работаете в частной компании и скажете: «Я провел A/B-тестирование, в результате которого общий доход компании увеличился с 20 до 23 миллионов долларов», то это плохой ответ, поскольку вы раскрываете конфиденциальные сведения о финансовом положении.

Возможно, вы ищете вакансии, на которых предстоит использовать другие технологии, будь то облачные провайдеры, диалекты SQL или основные языки программирования. В этом случае пользуйтесь теми же стратегиями, которые вы применяли, когда описывали свой опыт работы с точки зрения переносимых навыков. Предположим, вы работали с R, а компания использует Python. Можете сказать что-то вроде: «Я знаю, что мне потребуется немного времени, чтобы

освоить синтаксис Python, но я уже начал делать это на онлайн-курсах. Зато за четыре года программирования на R я разработал веб-приложения, создал пакеты и проанализировал большие наборы данных; благодаря этим навыкам я быстро стану сильным программистом на Python».

В этой главе мы уже упоминали синдром самозванца: помните о нем при подготовке к интервью. Когда вы только получили диплом и ищете первую работу или же начинаете новую карьеру в Data Science, можно легко сказать: «Я пока этого не знаю». (По крайней мере, это легко, если вы себя в этом убедите.) Но, когда у вас уже есть немного опыта, вы можете чувствовать себя неловко, если чего-то не знаете. Если на интервью произошло именно так, не бойтесь сообщить об этом. Можно сказать, что вы не нашли возможности применять эту технологию или хотели бы узнать о ней больше, но пока она не была частью вашей работы. Предположим, вас спрашивают об алгоритмах МО, но вы работали над статистическим моделированием, SQL, очисткой данных, а также общались со стейкхолдерами, а машинным обучением занимались инженеры. Никто не знает всего, и мы надеемся, что до сих пор вы неплохо справлялись со своими обязанностями; верьте в себя. Покажите проделанную работу; если вы изучали определенную тему раньше, то сможете освоиться быстрее, даже если прежде не использовали эти знания. Всегда лучше продемонстрировать готовность учиться, чем пытаться соврать о своей компетенции.

15.3. Поиск новой работы для трудоустроенных

Если ваш путь к должности дата-сайентиста лежал через буткемп, то вы, скорее всего, искали ее, будучи безработным. Если вы учились в университете, всем было понятно, что вы берете отгулы для прохождений интервью, на подготовку резюме или сопроводительного письма (см. главу 6). Но если вы работаете полный день, ваш руководитель вряд ли обрадуется, услышав, что вам нужно несколько выходных для поиска нового места. Так как же найти для этого время?

То, чем можно заняться когда угодно, например обновить резюме и сопроводительное письмо, изучить вакансии и откликнуться на них, сделать тестовые задания, стоит делать в свое свободное время. Не нужно это афишировать, да и пока что вы обязаны своей нынешней компании продолжать хорошо выполнять свои обязанности. Но интервью почти всегда проходят в обычное рабочее время. Если они проходят по телефону, мы рекомендуем ответить на звонок в переговорной, конференц-зале или в другом месте, где вас не услышат.

Однако более поздние этапы интервью, как правило, проводятся в офисе. Если они длятся час или два и проходят недалеко от вашего места работы, можете сказать начальству, что вам нужно на прием к врачу. Если дольше и ваша компа-

ния позволяет вам работать из дома, можете пойти на интервью в рабочее время и отработать только несколько часов (или попытаться отработать позже), но вы должны быть уверены, что в этот день нынешний работодатель не ждет вашего звонка или быстрого ответа. Можете также попробовать назначить интервью ближе к концу дня и поработать утром.

Безусловно, запланировать интервью легче, если вы ищете работу в своем городе. Но если вы собираетесь переехать, большинство компаний приглашают вас на личную встречу в будний день. Тогда вам едва ли удастся избежать отгула на целый день; большинство людей в таких ситуациях обычно «заболевают». Но как вы понимаете, если у вас много финальных интервью, то «болеть» становится весьма сложно.

Это одна из причин, по которой мы рекомендуем откликаться на вакансии стратегически. Если у вас будет больше десяти телефонных интервью и два в офисе, это наверняка заметят, а ваша работоспособность явно пострадает. Будьте избирательны: в первую очередь откликайтесь, а подготовкой занимайтесь после первого звонка. Если вы работаете в стартапе и хотите перейти в более крупную компанию, не откликайтесь на вакансии в других стартапах, даже если описания должностей выглядят великолепно. Если во время телефонного разговора выясняется, что эта должность больше связана с инженерией данных, а вы хотите заниматься анализом, можно смело завершать разговор, даже если интервьюер хочет его продолжить.

Позиции, от которых вы бы все равно отказались, можно использовать для практики, но не переусердствуйте. Опытные дата-сайентисты очень востребованы, а это значит, что как только вы объявите, что ищете новую работу, то вызовете большой интерес со стороны рекрутеров и менеджеров. Замечательное чувство: вы нравитесь людям! Тем не менее не позволяйте себе тратить время на интервью с компанией, которая вам не подходит, о чем вы заранее знаете. Даже если такое внимание льстит, это непродуктивная трата времени.

Когда вы ищете новую работу, легко упустить текущую. Чтобы чувствовать себя хорошо при мысли о смене места, вы часто вспоминаете о вещах, которые не нравится здесь и сейчас, а это может подорвать мотивацию. Все же постарайтесь работать на совесть; однажды вам может потребоваться рекомендация руководителя; кроме того, вы пока что получаете здесь зарплату.

Возможно, во время поиска вы придете к мысли, что на другой стороне трава не всегда зеленее. Другими словами, вы не можете найти интересных альтернатив в плане обязанностей: или все компании предлагают зарплату поменьше и соцпакет похуже, или же на новом месте не будет той гибкости, которой вы пользуетесь сейчас. Отказаться от поиска совершенно нормально! Оценка текущей работы — не зря потраченное время. Если вы передумаете увольняться, мы рекомендуем вернуться к совету в разделе 15.1.2: убедитесь, что вы попытались решить все возможные проблемы на нынешнем месте.

Возвращение в университет

Поработав в Data Science, вы можете решить, что хотите вернуться в университет и получить более формальное университетское образование — очно либо заочно. Если вы думаете об этом, мы рекомендуем вам вернуться к главе 3, где даем советы по поиску хорошей программы.

Тщательно продумайте, окупятся ли ваши затраты времени и денег, учитывая, что вы уже доказали способность работать дата-сайентистом. Некоторые причины, по которым возвращение в университет имеет смысл: вы решили, что хотите заниматься исследовательской работой, для которой нужна кандидатская степень; вы получили обратную связь от компаний, в которых хотите работать, и вам нужна степень магистра (вы ее не просто увидели в описании вакансии) или вы обнаружили, что для прогресса нужны определенные навыки (например, глубокое знание алгоритмов), а бесплатные онлайн-варианты вам не подходят.

Если вы решите учиться очно, то можете более открыто говорить об этом руководителю, чем в случае с переходом на новую работу. Если вы работаете в более крупной фирме, то компания может даже оплатить часть обучения, если оно заочное, а вы продолжите работать полный день или согласитесь вернуться после получения диплома. Даже если нет, ваш руководитель мог бы написать отличное рекомендательное письмо. Хороший менеджер знает, что университет предлагает нечто совершенно иное, чем работа, и должен поддерживать ваш выбор.

15.4. Сообщение об увольнении

Если вы решили уволиться и приняли предложение другой компании, сообщите об этом руководителю. Как правило, вы должны уведомить об этом минимум за две недели, если ситуация не является катастрофической. Маловероятно, но вполне возможно, что, как только вы сообщите об увольнении, руководитель скажет, что сегодня ваш последний рабочий день. Вы должны быть готовы к такому исходу и убедиться, что сохранили себе что-нибудь важное с рабочего компьютера.

Ваш босс должен быть первым, кто узнает о вашем уходе. Запланируйте встречу (назовите ее «Обсуждение карьеры», а не «Через две недели я ухожу») или поговорите с ним с глазу на глаз в нерабочее время. Вы должны сообщить об этом либо лично, если работаете в одном офисе, либо по телефону или по видеосвязи, если трудитесь удаленно; уведомление по имейлу — дурной тон. Для начала поблагодарите за то, что вам помогли в карьере, и за возможности, которые вам предоставили. Заверьте руководителя, что сделаете все возможное, чтобы ваш уход прошел максимально безболезненно для компании; вы можете, например, предложить несколько идей, прокомментировать код или предложить кому-то взять на себя часть вашей работы, но лучше обсудите это с менеджером. Вполне нормально беспокоиться о том, как сообщить руководителю об увольнении, но помните, что смена работы — естественная часть карьерного пути.

15.4.1. Рассмотрение контроффера

Возможно, руководитель попытается убедить вас остаться, сделав контрпредложение, поскольку нанимать нового человека дорого и рискованно. Он может попросить вас встретиться с его начальником, который имеет право дать вам повышение, дополнительные опционы, разовый бонус, ускоренную оценку эффективности или другие стимулы, чтобы вы остались.

Мнения о том, стоит ли принимать контроффер от нынешней компании, расходятся. С одной стороны, теперь компания знает, что вы «склонны к побегу», и может впредь не доверять вам серьезные задачи. Ситуация также может осложнить взаимоотношения с руководителем. С другой стороны, компания может захотеть устранить основную причину вашего ухода. Это могут быть как финансовые вопросы, так и перевод в другую команду.

Мы надеемся, что убедили вас в важности открыто общаться с руководителем. Если вы поговорили с руководителем и по-прежнему хотите перейти на новое место, вряд ли желаемые изменения появятся в контроффере. Мы не поддерживаем идею смены работы в качестве крайней меры и считаем, что следует принять решение об увольнении задолго до возникновения критической ситуации. Мы также настаиваем на том, что важно говорить о своих желаниях до того, как начнет расти ваше недовольство. Нужно понимать, что если вы дотянули до последнего, то вряд ли что-либо изменится в лучшую сторону, даже если вы поговорили с руководителем.

Начальник может попытаться подчеркнуть вашу ценность для команды и то, насколько сложно будет ее членам, если вы уйдете. Это может вызвать чувство вины, особенно если эти люди в целом хорошо к вам относились. Но помните, что вы никого не предаете. В конце концов, работа есть работа и, несмотря на слоганы некоторых стартапов, компания — это не ваша семья. Вы всегда должны относиться к коллегам с уважением и стараться работать хорошо, но вы не обязаны оставаться в компании бесконечно. И вообще: вы просто уходите из компании, а не умираете! Если вы сблизились со своими коллегами, то все еще можете встречаться с ними где-нибудь и, возможно, однажды даже снова поработать вместе.

15.4.2. Как сказать команде

Спросите руководителя, как бы он хотел сообщить эту новость остальной команде. Вас могут попросить подождать несколько дней, пока будет составлен план на переходный период, а затем руководитель представит его команде, а вы скажете, что увольняетесь. Вас могут спросить, хотите ли вы рассказать об этом всем на обычном собрании или лично. Размышляя о том, как сообщить эту новость, учитывайте размер команды. Если многие годы вы работали с одними и теми

же пятью людьми, то поделиться можно с каждым из них. Если же вы работаете в команде из 20 человек и с десятком стейкхолдеров, то легко ощутить эмоциональное истощение, объясняя все лично каждому по полчаса.

Одна из ошибок, которых следует избегать, — планирование встреч с коллегами до разговора с начальством, даже если они пройдут после обсуждения новости с руководителем. Если коллеги начнут подозревать, почему вы вдруг захотели с ними встретиться, и спросят, не потому ли, что вы увольняетесь, это будет действительно неловкая ситуация: вам придется либо солгать, либо рассказать все до того, как об этом узнает руководитель.

Чек-лист перед увольнением

Прежде чем отправиться в новый путь, проверьте, что у вас есть следующее:

- Контакты отдела кадров на случай, если что-то понадобится, например вы захотите узнать о своих опционах.
- Любые личные изображения, пароли или файлы, которые есть только на рабочем компьютере.
- Информация для входа на портал льгот и выплат в виде акций.
- Копии трудовых соглашений, писем с офферами и соглашений о расторжении договора.
- Информация о том, как будут оплачены оставшиеся дни отпуска.
- Варианты продления медстраховки (если вы не выходите на новую работу сразу).
- Если вы внесли средства на накопительный счет на медицинское обслуживание или на уход за зависимым лицом, то уточните, в какой срок можно их реализовать (как правило, это последний день работы или последний день последнего месяца трудоустройства). Эти средства нужно потратить, либо они сгорят.

Большинство коллег спросит, почему вы уходите. Убедитесь, что вам есть что ответить, и постарайтесь говорить в положительном ключе, сосредоточившись на новых возможностях и на том, за что вы благодарны нынешней компании. Даже если вы подружились с коллегой, все равно остерегайтесь плохих отзывов о нынешнем работодателе. Помните, что, возможно, вы захотите вернуться и вам не нужна репутация человека, который наплевал в свой же колодец. У некоторых складываются очень близкие взаимоотношения с руководителем, но даже в таком случае не стоит слишком много говорить о негативных аспектах текущей работы, поскольку это может повредить вашей дружбе. Помимо всего прочего, поддержание хороших взаимоотношений с бывшими коллегами и руководителем может иметь неоценимое значение для вашей дальнейшей карьеры, поскольку вы еще можете с ними встретиться или просить рекомендации.

Уходя из компании, оставьте коллегам возможность связаться с вами (по электронной почте, LinkedIn, Twitter и так далее). Коллеги будут рады оставаться на связи; кроме того, это хороший способ быть частью функциональной сети контактов.

15.4.3. Упрощение передачи дел

Лучший способ оставить хорошее впечатление — передать дела как следует. Возможно, вам не удастся найти себе замену, но вы можете подготовить команду, пока будут искать человека на ваше место (если захотят). Составьте для руководителя документ о передаче дел, перечислите свои обязанности, укажите, какие задачи вы можете завершить сами, а что нужно передать коллегам (предложите, кто мог бы с ними справиться), а также укажите, какие задачи придется отложить для нового сотрудника. Помимо передачи дел человеку, который берет на себя ваш проект, возможно, потребуется представить его внешним партнерам или клиентам, а также сообщить им, что вы больше не будете работать над этими задачами.

Постарайтесь подобрать все хвосты. Если у вас есть работа, которая потенциально пригодится другим, но хранится только на вашем компьютере, добавьте ее в репозиторий git или поделитесь Гугл-документом. Скорее всего, вас не будут сильно нагружать задачами в последние несколько недель, потому что все знают о вашем уходе. Это дает вам время доделать то, что вы могли откладывать на потом, например задокументировать все созданные вами процессы. Вот список того, что еще можно сделать:

- *Составьте инструкции*. Вы когда-нибудь были «палочкой-выручалочкой», человеком, который отлично разбирался в том, как организованы финансовые данные или передовые методы A/B-тестирования? Вы ничем не замените свое присутствие, но, оставив презентации, внутренние сообщения или документацию, поможете заполнить некоторые пробелы, которые появятся после вашего ухода.
- *Организация файлов*. Даже если вы добавите все на GitHub, это никому не поможет, если «все» — это 100 файлов с такими именами, как random_stuff и misc_analyses. Постарайтесь упростить навигацию и добавьте пояснения, если они нужны.
- *Добавление комментариев и пояснений к анализу*. В идеале любой эффективный анализ предполагает, что вы уже прокомментировали выводы по коду. Если вы не успели завершить некоторые дела и думаете, что кто-то продолжит эту работу, то можете дополнить комментарии. Не обязательно комментировать каждый фрагмент кода, но будет хорошо, если вы укажете какие-нибудь особенности данных (и то, как вы с ними работали), опробованные варианты и причины, по которым вы выбрали те или иные аналитические методы.

Худшее, что вы можете сделать, — это забыть, что вы единственный человек в компании, знающий, как решать определенную задачу, и никому ее не передать. Если вы так сделаете, то будете получать гневные звонки и имейлы с вопросами о том, как выполнять эту работу, в то время как будете пытаться освоиться на новом месте. Забыть, что вы единственный, кто знает пароль к определенной системе, — хороший способ нарваться на неприятности даже после увольнения. Некоторые работодатели не умеют расставаться и могут названивать по поводу ваших прошлых проектов. Вам же лучше, если у вас будет возможность отсылать таких бывших коллег к документации, которую вы подготовили перед увольнением, до тех пор пока человек не поймет, что вы больше не работаете в компании. Если вы оставите после себя бардак и неразбериху, то вряд ли сможете воспользоваться сетью полезных контактов, которую создали в этой компании.

Мы надеемся, после прочтения этой главы вы поняли, что, несмотря на сомнения в своем решении уволиться и на сопутствующий стресс, это нормальный процесс и всегда есть способы сгладить острые углы. Мы уже много раз говорили, что совсем не многие ваши решения окончательны. Сам факт поиска новой работы не означает, что обязательно нужно увольняться, и даже если вы уйдете из компании, то сможете вернуться туда через несколько лет. Сосредоточьтесь на том, чтобы работа максимально соответствовала вашим целям в карьере.

15.5. Интервью с Амандой Касари, техническим менеджером Google

Аманда Касари (Amanda Casari) — технический руководитель в Google, член команды по взаимодействию с разработчиками облачных вычислений Google. Ранее она была ведущим руководителем продукта и дата-сайентистом в SAP Concur. Она также пять лет прослужила в BMC США и получила степень магистра электротехники.

Как вы понимаете, что пора искать новую работу?

Мой совет — понять, чем бы вы хотели заниматься, сможете ли вы делать это на текущей должности, а также определить, на каком этапе развития находятся продукт, команда и компания. Что касается меня, я лучше всего себя чувствую в условиях, когда все постоянно меняется. Мне нравится работать над проектами в самом начале, во время формирования идей, а также на завершающих этапах разработки. Если бы мне пришлось тратить бóльшую часть времени на оптимизацию моделей для одноразрядного увеличения процента или на настройку гиперпараметров, я работала бы не очень эффективно. Еще я обращаю внимание на уровень слаженности команды. Хотите ли вы присоединиться к группе специалистов, в которой уже сформировались устоявшиеся взаимоотношения

и культура, или к той, которая только сложилась? В целом мое место там, где продукт находится на указанных этапах развития и где команда формируется и не имеет какой-либо устоявшейся культуры. Эти факторы играют решающую роль при оценке, устраивает ли меня текущая должность или пора искать что-то более перспективное.

Случалось ли так, что вы начинали искать работу, но в итоге решали остаться?

Постоянно. Когда я просматриваю другие вакансии, то могу понять, чем на самом деле могла бы заниматься в нынешней компании. Это своего рода удовольствие — найти возможности, а не ждать, пока мне их кто-то предоставит. Это часть моей более широкой философии, согласно которой текущие обязанности должны формироваться в процессе общения с руководителем. Со своими инженерами я стараюсь вести открытые и честные диалоги, в ходе которых они говорят мне, чем хотели бы заниматься. Зная это, я могу выяснить, есть ли подходящая работа в текущей команде; если нет, то сможем ли мы найти какой-нибудь сторонний проект с 20 %-ной занятостью. Таким образом, человек может разобраться и понять, что ему действительно по душе.

Знаете ли вы людей, которые слишком долго работают на одном месте?

Конечно, да. Я наблюдала, как у некоторых возникает своего рода комплекс героя: им кажется, что никто другой не сможет справиться с их работой. Правда в том, что никто не будет делать эту работу точно так же, но это не значит, что ее вообще не удастся выполнить. Иногда кто-то работающий в команде очень давно может вредить работе, потому что он помнит все проблемы и решения, которые когда-либо принимала компания. Он может сказать: «Мы проверяли эту идею два года назад, но она не сработала, поэтому не стоит ее рассматривать». Но команду это может тормозить, поскольку она будет сосредоточена не на использовании имеющихся возможностей, а на решениях, применявшихся раньше.

Я также видела людей, которые изрядно устали от руководителей и тратят много времени просто на выражение своего недовольства. Они всегда готовы поделиться сплетнями о положении дел, а для компании это плохо. Вам не нужны токсичные сотрудники, которые будут заражать коллег своим недовольством.

Наконец, я видела людей, которые не стремятся развиваться на текущей должности: они просто делают то, что от них требуют, и не более того. Это может быть хорошо, когда вы новичок, но от опытных сотрудников и руководителей я ожидаю чего-то большего. Я хочу, чтобы опытные сотрудники стремились применять свои способности на более высоком уровне. Если перед вами стоит задача, вам следует подумать, как найти масштабируемое, воспроизводимое решение, которое позволит справиться с классом схожих задач, а не только с одной конкретной.

Можно ли слишком часто менять работу?

Рассматривая соискателей, я могу спросить, работал ли кто-то из них на должности меньше года. По стажу работы рекрутер пытается понять, не уйдет ли человек через несколько месяцев, ведь наем сотрудника и его адаптация — долгий и дорогостоящий процесс. Однако при том что в других отраслях люди могут работать на одной должности как минимум два-три года, я считаю, что для ИТ-сферы это слишком долго. Здесь два-три года — это несколько проектов, выполненных несколько раз, поэтому увольнение раньше этого срока я считаю вполне обоснованным. На самом деле даже года вполне достаточно. Если вы хотите уйти раньше, чем через год, из-за психического или эмоционального состояния, это тоже приемлемая причина: ни одна работа не стоит того, чтобы рисковать здоровьем.

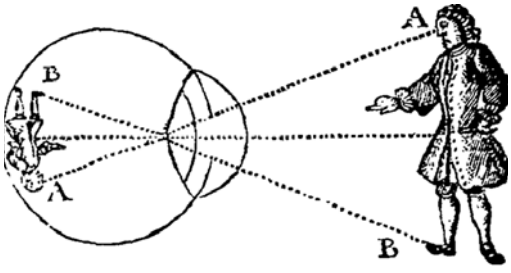
Ваш последний совет для начинающих дата-сайентистов.

Найдите единомышленников и людей, которые могут вам помочь. Я многое приобрела благодаря близкому другу, который рассказывал мне о способах общения и вариантах трудоустройства. То, что рядом оказался человек с таким опытом, было бесценным и действительно помогло мне осознать свою важность и чувствовать себя уверенно при переходе на новую должность. Сообществ, где вы можете найти единомышленников, много, но не каждое из них может дать вам ощущение причастности. Не стоит оставаться там, где вам некомфортно, будь то язык, люди, которые там собрались, или цели сообщества. Если же у вас не получилось найти комфортное пространство, поищите кого-то, на чьем месте вы хотели бы оказаться в подобной группе, и спросите, может ли он помочь вам такую группу сформировать.

Итоги

- Принимая решение о поиске нового места, задайте себе четыре вопроса: 1) учитесь ли вы на работе чему-то новому; 2) просили ли вы начальство о смене обязанностей; 3) совпадают ли карьерные цели у вас и у руководителя; 4) думали ли вы о том, чего хотите (и не хотите) на следующем месте работы.
- Многие принципы, которые вы применяли, когда искали работу в Data Science в первый раз (и о которых мы говорили во второй части этой книги), продолжают работать и позже, но в дальнейшем вам также следует подумать о преимуществах и недостатках своей первой работы. Будьте готовы рассказать о своем опыте в позитивном ключе, соблюдайте конфиденциальность, спланируйте, как совмещать интервью со основной работой.
- Сообщите руководителю об увольнении за две недели, а затем постарайтесь максимально передать дела команде, подберите хвосты, задокументируйте все, о чем сейчас знаете только вы, поделитесь любым полезным кодом.

16



Вверх по карьерной лестнице

В этой главе

- Варианты развития после должности старшего дата-сайентиста.
- Возможности и риски потенциальных карьерных траекторий.

В последней части этой книги мы говорили о том, как построить карьеру: научиться переживать неудачи, стать членом сообщества и сменить работу. По мере роста вы в конечном итоге захотите решить, в каком направлении будете развиваться дальше. Не всегда очевидно, какие варианты будут в дальнейшем у дата-сайентистов; стать руководителем — одна из возможностей, но она не единственная.

В этой главе мы рассмотрим три общие карьерные траектории для специалиста по данным: стать начальником, техническим руководителем или техлидом и перейти в независимый консалтинг. Мы также расскажем о преимуществах и недостатках каждого из этих вариантов.

Руководящая должность — именно то, что представляет большинство людей при мысли о карьерном росте. *Руководители* — это люди, которые возглавляют команды, а еще отвечают за наем новых сотрудников и продвижение по службе, определение стратегии и наставничество по вопросам карьеры. *Ведущие дата-сайентисты* — это настоящие мастера своего дела, и компании доверяют им решение сложных технических задач. *Независимые консультанты* — это специалисты по данным, у которых достаточно знаний и контактов, чтобы зарабатывать на

жизнь фрилансом. На рис. 16.1 представлены общие варианты развития карьеры на продвинутом уровне.

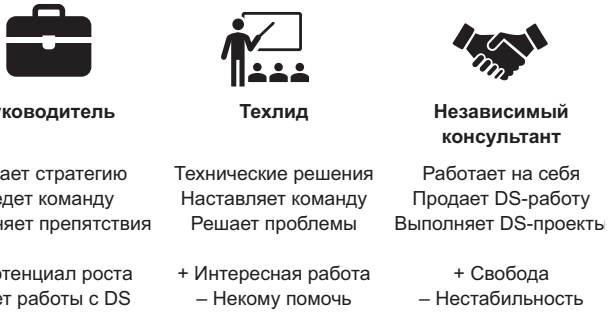


Рис. 16.1. Пути развития, о которых пойдет речь в этой главе

Знайте свой уровень

На работе вы приобретете ценные навыки (о которых говорилось в предыдущих 15 главах). Но по мере роста вашего набора навыков и профессиональной зрелости в какой-то момент вы перестанете быть джуниором. Невозможно знать наверняка, когда это произойдет, поэтому сложно выбрать подходящий момент, чтобы попросить о повышении или о новой должности. Имейте в виду, что у каждой компании свои уровень и ожидания, поэтому одно и то же название специалиста может означать совершенно разные обязанности в двух разных организациях. Внутри компании может даже существовать матрица навыков, точно описывающая, чем уровни отличаются между собой, но ее можно интерпретировать по-разному: двусмысленность есть всегда. Чтобы вам было легче, ниже мы даем подробное руководство по разным ожиданиям к уровню дата-сайентиста:

- *Джуниор*. Человек, который может выполнить DS-задачу, получив четкое указание. Если джуниору скажут использовать алгоритм кластеризации для сегментации новых клиентов по покупательским качествам, он сможет сделать это под управлением руководителя. Если возникают технические проблемы, например ошибки в коде или невозможность подключить системы данных, для решения ему может потребоваться консультация с другими членами команды.
- *Синьор*. Человек, который может не только выполнять поставленные DS-задачи, но и определять, что еще необходимо решить. Он не только смог бы выполнить приведенную в качестве примера сегментацию посредством кластеризации (описанную выше), но также понимал бы, что аналогичный алгоритм можно использовать и для существующих клиентов (а затем выполнить то, что требуется). Он умеет решать технические задачи и приходит на помощь, если у других возникают трудности.
- *Кто-то выше синьора*. На этом уровне задачи в основном связаны с помощью другим специалистам. Таким образом, человек обычно перестает быть синьором, когда постоянно помогает другим в решении задач, разрабатывает стратегии и видит более широкую картину.

По мере развития карьеры вам нужно будет сосредоточиться на одном из этих путей. Имея четкую цель, вы с большей вероятностью добьетесь желаемого. При этом может быть и так, что чем ближе вы подходите к реализации цели, тем сильнее чувствуете, что это не то, чего вы хотите. К счастью, вы всегда можете все переиграть. Стать руководителем и понять, что это не ваше? Нормально. Перейти из индустрии в консалтинг, а затем вернуться обратно? Бывает. Иногда сложно менять принятые решения, но учиться на своих ошибках — это быстрый способ расти как личность!

16.1. Путь руководителя

Дата-сайентисту может казаться, что продвижение на должность руководителя — это естественный шаг в карьере: ведь у каждого есть начальник, перед которым нужно отчитываться. Однако повседневные задачи руководителя не всегда выглядят так, как мы их себе представляем.

Руководитель — это человек, который отвечает за команду людей, успешно выполняющих свои задачи. Он обычно (но не всегда) делает следующее:

- *Определяет задачи для команды*. Это можно делать на стратегическом уровне, например решать, за какие крупные проекты следует взяться. Или на тактическом: решать, какие функции должны быть включены в продукт.
- *Определяет, кто должен быть в команде*. По согласованию с отделом кадров и другими сторонами руководитель обычно выбирает, кого взять в команду, а кого отпустить. Он координирует процесс интервью и принимает взвешенные решения.
- *Берет на себя наставничество*. Каждый член команды сталкивается с уникальными задачами, которые нужно решить, и руководитель в этом помогает. Руководитель регулярно проверяет каждого сотрудника и дает советы и рекомендации, которые помогут им в решении задач.
- *Решает сложные вопросы, с которыми столкнулась команда*. Если что-то мешает продуктивной работе (например, другая команда не желает предоставить необходимый доступ к данным), задача руководителя — найти решение.
- *Управление проектами*. Руководитель должен мониторить работу команды и следить за соблюдением графика. Хотя во многих отделах есть менеджер проекта, специально назначенный для этой задачи, руководитель все равно должен держать руку на пульсе.

Вместе эти задачи охватывают широкий спектр работы и предполагают высокую ответственность. Руководитель должен постоянно общаться с людьми внутри и за пределами своей команды. Ему нужно знать, как идут дела у коллег,

что происходит с проектами и чего ждать в ближайшем будущем. Эти задачи составляют огромный объем работы, и, если какая-либо из них не будет выполнена хорошо, пострадает вся команда.

Важно отметить, что в основном перед руководителем стоят нетехнические задачи; обычно он не создает модели машинного обучения и не проводит анализ, на основании которого компания будет принимать решения. У руководителя нет на это времени, а даже если бы оно было, все равно такими вещами лучше заниматься его подчиненному. Чтобы стать руководителем, придется отказаться от большей части того, ради чего вы стали дата-сайентистом: от желания использовать данные для решения интересных задач. Вместо этого вы будете помогать специалистам, которые будут выполнять эту работу.

16.1.1. Преимущества работы руководителем

У работы руководителя есть множество преимуществ. Во-первых, если вы человек, который ненавидит, когда другие делают глупые, на ваш взгляд, вещи, у вас есть все права и возможности, чтобы повлиять на ситуацию и сократить масштаб бесполезной работы. Например, если вы считаете, что определенная модель машинного обучения принесет компании пользу, то можете попросить команду разработать ее, но если вам это кажется плохой идеей, то вы должны убедиться, что команде не поручат ею заниматься. Это невероятно приятно — выбирать направление, а затем видеть, как команда добивается успеха. Иногда другие руководители требуют или настоятельно просят вашу команду что-то сделать, и, хотя у вас нет полного контроля, вы имеете право голоса.

Рост до должности руководителя обеспечивает повышение зарплаты. Перед вами также открываются новые двери: роль старшего менеджера, директора или вице-президента. Каждая из них предполагает более высокую оплату и более широкие возможности управления в компании. Вы даже можете дорасти до уровня, на котором будете курировать и другие сферы деятельности (например, изучение клиентских требований или разработку ПО), а не только анализ данных. В конечном итоге вы можете полностью отказаться от Data Science и возглавить другие области.

Если вам нравится учить других и помогать им, должность руководителя отлично подойдет. В основном вы будете помогать специалистам в команде добиваться успеха, работать с ними, обучать их тому, что знаете сами, и помогать им в сложных ситуациях. Хороший руководитель подобен бизнес-терапевту: он общается с человеком примерно час, а затем помогает ему справиться с проблемами.

Наконец, как руководитель вы можете иметь огромную сферу влияния. Быть человеком, который принимает окончательное решение о выпуске нового продукта или открытии филиала в другой стране, действительно круто! По мере

вашего роста увеличивается и ваше влияние. При выборе этого пути у вас есть возможность когда-нибудь управлять компанией.

16.1.2. Недостатки должности руководителя

Самый большой недостаток должности руководителя заключается в отсутствии времени для занятия Data Science. Вы будете обсуждать анализ данных, наставлять специалистов и размышлять о стратегиях. Ваш день может состоять из 30-минутных совещаний на совершенно разные темы: от выбора командной стратегии и до общения со стейкхолдерами для получения финансирования, включая беседы тет-а-тет с неэффективными сотрудниками.

В том, чтобы перестать заниматься Data Science в качестве своих обязанностей, есть два основных недостатка:

- Вы стали дата-сайентистом, потому что вам нравится работать с данными, так что вы бросаете работу, ради которой так долго учились.
- Если вы не занимаетесь данными, у вас нет практики, вы перестаете следить за последними тенденциями. Если вы решите, что должность руководителя — это не ваше, и захотите снова стать исполнителем, может обнаружиться, что ваши навыки Data Science заржавели.

Еще один недостаток — вы все еще ограничены вышестоящим руководством. У вас могут быть отличные стратегические идеи, а ваш начальник их не поддержит. Кроме того, вы вынуждены вести команду по пути, выбранному вашим начальником, даже если вы с ним не согласны. Необходимость сохранять позитивное отношение к работе, с которой вы не согласны, ради блага команды может очень раздражать. Если подчиненные почувствуют это раздражение, вы снизите их удовлетворенность работой.

На руководящей должности у вас появляется гораздо больше забот. Вы должны заботиться о своей собственной эффективности, так же как тогда, когда были исполнителем, но теперь вам нужно беспокоиться еще и о производительности остальных участников команды. Вы должны заботиться о том, насколько они удовлетворены своей карьерой, а еще о том, что происходит на вышестоящих уровнях, и быть в курсе политики компании. Вам придется переживать, будет ли финансирование у команды и не будет ли отменен проект, который продвигается слишком медленно. Наличие такого количества забот, большинство из которых находятся за пределами вашего непосредственного контроля, может стать источником колоссального стресса для людей определенного типа личности. Если вам сложно не брать работу домой, возможно, руководство не для вас.

Наконец, управление людьми требует совершенно иных навыков, нежели работа дата-сайентистом. Если вы решили заняться руководством, вам при-

дется освоить эти навыки и снова почувствовать себя не особо компетентным. Превращение из специалиста, который отлично справляется со своей работой, в новичка может вызвать стресс и заставить почувствовать себя несчастным. Хотя со временем вы всему научитесь, путь к статусу хорошего руководителя долог.

16.1.3. Как стать руководителем

Если вы рядовой сотрудник, но хотите стать руководителем, вам необходимо найти возможность для развития и тренировки лидерских навыков. Эти навыки включают в себя умение работать с людьми как младше, так и старше вас, способность видеть общую картину и навыки управления проектом.

К сожалению, не существует одного такого курса, на котором можно всему этому научиться; лучшее, что вы можете сделать, — это использовать ситуации, возникающие на вашей текущей работе, которые позволили бы развивать их. Скажем, когда вы можете взять на себя ответственность и инициативу за выполнение небольшого задания, порученного вашей команде, например настройки нового стека ПО или координации развертывания новой модели. Самое важное здесь — это управление процессом и принятие решений. Сначала это может показаться вам крайне неестественным; это совершенно нормальное ощущение, и оно пройдет. Книги по менеджменту и бизнесу могут быть в некоторой степени полезны, но только в том случае, если у вас есть возможность применять на практике полученные с их помощью знания.

Когда вы почувствуете, что достаточно развили навыки руководителя, можете переходить к следующему шагу — к поиску подходящей должности.

ПОЛУЧЕНИЕ ДОЛЖНОСТИ В СВОЕЙ КОМПАНИИ

Зачастую самый простой способ занять руководящую должность — продвигаться внутри своей компании. Самым простым он считается, поскольку те, кто будет утверждать вашу кандидатуру, — это люди, которые наблюдали за вашей работой и видели, как развиваются ваши навыки. Трудность этого пути заключается в том, что компании для начала должен потребоваться новый руководитель: либо один из действующих начальников должен уволиться или получить повышение, либо должно появиться вакантное место в соответствующей команде. В зависимости от компании такие ситуации возникают редко или крайне редко.

СОБСТВЕННАЯ КОМАНДА

Другой путь — стать руководителем, вырастив команду самостоятельно. Вы можете в своей нынешней компании начать проект, который в итоге потребует участия большего количества людей, или первым получить должность дата-сайен-

тиста в другой компании и затем создать там свою команду. Этот путь может быть чрезвычайно эффективным, потому что, выращивая команду самостоятельно, вы точно знаете, кто в ней участвует и как она функционирует. Работа требует, чтобы вы были в нужном месте в нужное время и были достаточно сильным лидером, чтобы быстро развивать всю инфраструктуру команды. Иногда людей, использующих такой подход, называют играющими тренерами, поскольку они одновременно являются первыми участниками команды и тренерами для коллег. К сожалению, возможность создать свою команду выпадает еще реже, чем получить ранее упомянутое повышение.

Роб Штамм (Rob Stamm), директор, курирующий команду AI @ T-Mobile: учиться управлять

Я многое узнал о том, что такое быть руководителем, когда первый раз поработал начальником на полную ставку. До этого я побывал продакт-менеджером, что означало руководить процессом его развития (но не людьми). Затем я получил должность старшего руководителя, и мне пришлось курировать нескольких разных менеджеров по продуктам и помогать им с их работой. Я с треском провалился на этой позиции. Я хотел быть старшим менеджером и возглавлять команду, но не переставал заниматься менеджментом продукта. Я не давал ребятам из своей команды выполнять свою работу: я делал это за них.

Через четыре месяца такой работы один из менеджеров продукта вошел в офис и сказал мне, что команда собирается уволиться из-за моих действий. Так я осознал одну важную вещь: нельзя одновременно быть и менеджером, и исполнителем. Тогда мне впервые указали на то, что руководитель должен давать возможность своей команде принимать решения, и это мне запомнилось.

В конце концов, я научился на этом опыте и начал управлять большими командами и крупными проектами. Быть лидером — невероятно благодарное дело: я помогаю своей команде стать лучше и чувствую удовлетворение, когда вижу результаты. Однажды в T-Mobile мы с командой ИИ запустили успешный проект, имея лишь идею и пару долларов в качестве финансирования. В нем я не занимался разработкой продукта и не писал код, а приходил на помощь коллегам, когда они где-то застревали или нуждались в пополнении финансовых или человеческих ресурсов. Это была по-своему плодотворная работа.

ПОЛУЧЕНИЕ ДОЛЖНОСТИ РУКОВОДИТЕЛЯ В НОВОЙ КОМПАНИИ

Последний способ — занять должность руководителя, освободившуюся в другой компании. Такой путь подразумевает, что вы способны продемонстрировать свои управленческие навыки, при том что ранее вы не занимали подобную должность. В сопроводительном письме и на интервью вам нужно будет рассказать обо всех проектах, которыми управляли, и о людях, которых наставляли, занимая позицию

рядового сотрудника. Вы сможете понять, есть ли у вас шанс использовать данный путь, по тому, какое внимание привлекает ваше резюме среди получивших его работодателей; компании отчаянно нуждаются в хороших руководителях DS-проектов, поэтому, если у вас хорошее резюме, на него обязательно отреагируют.

16.2. Путь ведущего дата-сайентиста

Ведущий дата-сайентист (или главный специалист по работе с данными, техлид, специалист по данным V, технический руководитель или другое название должности в зависимости от компании) — это человек в организации, который одновременно выступает экспертом в Data Science и помогает другим с техническими задачами. Для того чтобы стать руководителем, нужно все меньше и меньше заниматься Data Science, а у ведущих специалистов по данным все ровно наоборот. Но вместо того, чтобы заниматься этим в одиночку (хотя и так часто бывает), вам также будет поручено помогать другим специалистам своей области.

Техлиды обычно начинают свой путь с джуниора, затем переходят на должность синьора и продолжают расти. По мере развития карьеры они становятся опытными и более зрелыми в способности понимать проблемы и находить способы их решения. Вскоре они уже знают так много, что, если кто-то застрял, они могут вмешаться и быстро помочь. Другие специалисты компании подходят к ним и спрашивают, как решать проблемы и что сработает (а что нет).

Работа ведущего дата-сайентиста включает в себя несколько обязанностей:

- *Влияние на стратегии Data Science*. Ведущий специалист по данным должен разработать план решения DS-задач. Возможно ли построение модели для выявления мошенничества с платежами? Следует ли использовать нейронную сеть? Руководитель отвечает за идею и бизнес-план, ведущий специалист — за то, как это должно работать.
- *Наставничество джуниоров*. Поскольку техлид имеет большой опыт, он обязан делиться своими знаниями с младшими сотрудниками. Их рост так же важен, как и его непосредственная работа.
- *Поиск решений для сложных задач*. Когда группа специалистов по данным сталкивается со сложной технической задачей, ведущий дата-сайентист — это человек, который пытается разработать решение или сказать, что это невозможно.

В отличие от руководителя, техлид по-прежнему много занимается Data Science. Таким образом, эта роль отлично подойдет тем людям, которым нравится быть дата-сайентистами. Если все, о чем вы можете думать, — это Data Science, вы любите ходить на конференции, быть частью сообщества и узнавать больше о методах и техниках, то вы продолжите заниматься этим же в качестве техлида. Ведущий дата-сайентист является очень важным звеном команды, поэтому для

этой роли требуется человек зрелый, ответственный и достаточно опытный, чтобы при необходимости быстро расширять границы в новых направлениях.

Хотя работа техлида связана с данными, он все же редко выполняет ту же работу, что и рядовой исполнитель, например он не занимается отдельным анализом или созданием модели машинного обучения. Эти задачи требуют много времени и внимания, а ведущий дата-сайентист должен отвечать за множество проектов и направлений. За конкретные проекты обычно отвечают джуниоры и синьоры, а техлид их координирует.

Как просить о повышении

В какой-то момент вы почувствуете, что готовы перейти на более высокий уровень, но не получите повышения. Это может расстраивать (вы ведь готовы, так почему же не готовы остальные!?), но эта ситуация не безнадежна. Лучшее, что вы можете сделать, — это отстаивать свою кандидатуру. Сообщите своему руководителю, что заинтересованы в переходе на более высокую должность и хотите вместе разработать план действий. Если у вас понимающий начальник, он должен быть этим доволен: если вы недвусмысленно говорите о том, что готовы к изменениям для получения следующей должности, вы начинаете обсуждать варианты. Постарайтесь задать определенную цель и срок, когда вы хотите получить повышение. Например, это может быть следующая оценка эффективности. Цели должны быть как можно конкретнее, например «провести три технические презентации в компании» или «самостоятельно создать и развернуть весь API машинного обучения». Имея четкие цели с графиком, вы сможете наглядно видеть прогресс.

Если руководитель говорит, что вы пока не готовы, и объясняет причины, — прислушайтесь. Несмотря на то что отрицательная оценка всегда расстраивает, начальство считает, что вы еще не достигли уровня, необходимого для повышения. Если руководитель говорит, что, по его мнению, вы готовы и он замолвит за вас словечко, постарайтесь предоставить ему как можно больше документации о своей работе и причинах, по которым вы способны выполнять новую работу. На основании этой документации руководителю будет легче продвигать вас.

Если, несмотря на все старания, вы все равно не можете получить желаемое продвижение по службе, это может говорить о том, что пора переходить в другую компанию. Есть множество случаев, когда люди настолько привыкли видеть кого-то в определенной роли, что не хотят рисковать, давая этому человеку возможность попробовать себя в новой. Сменив компанию, вы начинаете работать с группой людей, у которых нет ожиданий относительно вас, а это дает больше перспектив.

16.2.1. Преимущества работы ведущим дата-сайентистом

Ведущий дата-сайентист часто получает самые интересные задачи. Если у команды есть идея совершенно нового и неизведанного подхода к анализу данных,

вам придется попробовать ее впервые. Если необходимо интегрировать сложный технический стек, вы обязательно будете в центре событий. Если проект должен быть запущен, но команда просто не может заставить модель работать, вы тоже будете заниматься этим вопросом. Быть в центре событий может приносить удовольствие. Это интересно с технической стороны, и вы почувствуете собственную ценность, когда будете снова и снова помогать коллегам выйти из сложного положения. Команда будет знать, что не стоит отвлекать вас ради простых, второстепенных проектов, а это означает, что вам придется заниматься ими меньше, чем большинству дата-сайентистов.

Ваш руководитель также будет понимать, насколько важно для вас идти в ногу с технологиями, а это означает, что компания будет закладывать на вас бюджет на посещение конференций и выделять время для экспериментов с новыми технологиями. Руководитель может пожелать, чтобы вы продвигали свою компанию во время выступлений на конференциях, а так как вы работали над интересными задачами и наверняка делали работу, которой стоит поделиться, вам есть что рассказать. Если вы попросите у начальства ресурсы, например деньги, на опробование нового облачного сервиса, обычно их выделяют без вопросов. Руководитель доверит вам самостоятельно распоряжаться своим временем и бюджетом эффективно и не тратить их напрасно: такая возможность дается не каждому дата-сайентисту.

Вы также будете постоянно говорить о Data Science. Как наставник вы будете рассказывать джуниорам о различных подходах, работать с ними и оттачивать их идеи, указывать области, в которых могут возникнуть сложности при анализе данных. Для человека, который любит Data Science, эта роль может быть довольно интересной.

Планирование работы с данными может давать много прав и возможностей. Будучи человеком, который решает, какие типы моделей использовать, как структурировать данные и как масштабировать проект, вы с большей вероятностью увидите, что реализация проектов идет по вашей задумке. А поскольку вы являетесь экспертом в Data Science, ваши проекты будут иметь много шансов на успех! Джуниорам не всегда удается самостоятельно разработать подход к анализу, реализовать его и увидеть результаты без участия других лиц в принятии ключевых решений.

16.2.2. Недостатки должности ведущего дата-сайентиста

Самая большая сложность в роли ведущего дата-сайентиста заключается в том, что вам не к кому обратиться за помощью, если вы застряли. У джунов обычно есть наставник или синьор, которому можно задать вопросы; порой они даже

могут просто загуглить решение. В вашей команде вряд ли будет кто-то выше вас. Задачи, с которыми вы столкнетесь, часто будут настолько необычными или уникальными, что ни один поисковик не даст ответа. Следовательно, вам нужно будет уметь работать без возможности обратиться за помощью, что может быть весьма тяжело.

Учтите, что вам придется решать не только самые интересные, но и самые неприятные задачи. Если, например, датасет хранится в виде терабайтов плохо отформатированных файлов в формате .csv на сервере, к которому никто не прикасался годами, и со схемой, которую никто не знает, вам предложат выяснить, как использовать эти данные. На самом деле эта задача неинтересна; скорее это минное поле, по которому никто не может пройти. Вы столкнетесь со множеством подобных задач, решение которых вы не сможете делегировать.

Обширные знания делают вас востребованным специалистом — вы будете очень заняты. Чаще всего у вас будет больше работы, чем времени на ее выполнение, и вам придется отказываться от интересных проектов, ведь вы не можете разорваться между ними. Переработки войдут в привычку, а чувство, что вы сделали недостаточно, станет вашим постоянным спутником. От вашей работы зависит много людей, поэтому работать только 40 часов в неделю или провести отпуск без ноутбука — нелегкая задача. Если вам нужна приятная работа без стресса, то роль ведущего дата-сайентиста не для вас.

16.2.3. Как стать ведущим дата-сайентистом

Если вы работаете специалистом по данным и продолжаете развиваться, то дорастете до ведущего. Это следующая ступень после должности синьора, ее расширенная версия. К сожалению, многим синьорам сложно перейти на следующий уровень. Чтобы стать техлидом, вы должны быть достаточно сильным специалистом, уметь руководить и эффективно работать независимо от других. Далее вам нужно привлечь внимание начальства к своим способностям и вкладу, который вы внесли в работу, а также найти тех, кто будет рекламировать вас как важнейшую часть команды. Обладая такой квалификацией, вы можете просить повышения.

Чтобы работать самостоятельно, вы должны уметь реализовать полный проект без руководства извне. Если руководитель поставил вам задачу проанализировать, где следует открыть очередной розничный магазин, он должен быть уверен, что вы справитесь без посторонней помощи. По мере вашего роста как специалиста способность к независимой работе должна стать чем-то естественным, поскольку вы набираетесь больше опыта. Старайтесь обращать внимание на моменты, где вы застреваете, и свои действия в этих ситуациях. Если вы попросите о помощи, то что получите такого, чего не смогли бы сделать самостоятельно? Чем больше задач вы будете решать сами, тем лучше.

По мере набора опыта постарайтесь обращать внимание на окружающих вас дата-сайентистов. С какими проблемами они сталкиваются? Можете ли вы им помочь? Если вы синьор, вполне вероятно, что младшие сотрудники имеют дело с задачами, которые вы уже решали раньше. Чем больше возможностей вы найдете для технического наставничества, тем больше будете помогать другим, а это отлично подходит для роли техлида.

Наконец, если у вас появляются новые идеи, ищите ситуации, в которых вы можете создать стратегии. Например, если компания хочет определить место для розничных магазинов, вам может прийти в голову идея использовать методы оптимизации местоположения. Идеи могут сработать или нет; если все получится, вы красавчик, а если нет, то будете лучше знать, как придумывать идеи. Всегда легко полагаться на других дата-сайентистов при выработке стратегии, но быть ведущим техлидом без этого навыка очень сложно.

16.3. Путь независимого консультанта

Многие мечтают быть самому себе начальником и иметь собственную компанию. В случае Data Science это обычно означает быть независимым консультантом: основать фирму, в которую другие будут обращаться за услугами по работе над специализированными DS-проектами. Теоретически организации должны нанимать стороннего консультанта только в том случае, если им нужен особый набор навыков для решения важной задачи. Если вы управляете собственной компанией, то можете оставлять весь доход себе, так что никакие деньги не достанутся другим жирным котам на руководящих должностях. Люди будут обращаться к вам, потому что знают, что вы отличный специалист по данным; вас будут ценить за компетентность.

Будучи независимым консультантом, вы должны объединять в себе функции целой компании, а это значит, что вам придется делать много разных вещей, например:

- *Продвигать свой бизнес*. Вы не сможете найти новых клиентов, если о вас не узнают. Посещение конференций, встречи со старыми коллегами или самореклама, например посты в блогах, — все это инструменты маркетинга.
- *Продавать*. Когда вы найдете компанию, которая заинтересована в том, чтобы нанять вас на DS-проект, вам нужно будет встретиться с ее представителями и сделать коммерческое предложение. Если вы не заинтересуете их, то не получите работу.
- *Реализовать проекты*. Это работа в области Data Science, для которой вас и наняли. Она также включает в себя управление проектами, предполагающее контроль процессов и решение любых сложных ситуаций, которые могут возникнуть в ходе работы (например, плохие данные).

- *Предоставить результаты*. Когда вы создали модель или провели анализ, вам нужно показать клиенту результат. Если ему все понравится, вы можете получить следующий проект, а если нет, то есть шанс потерять клиента.
- *Управлять компанией*. Компаниям необходимо платить налоги, готовить юридические документы, отслеживать счета и движение денежных средств, а также выполнять множество других мелких задач, которые со временем накапливаются.

Вам нужны навыки для выполнения всех этих задач, которые выходят далеко за рамки функций рядового дата-сайентиста.

В зависимости от типа клиентов, которые к вам обращаются, работа, скорее всего, будет наполовину связана с Data Science, а наполовину — с поддержанием бизнеса. Процесс будет протекать в определенном порядке: вы станете работать над одним проектом для компании-клиента, а к моменту его окончания будете готовиться к заключению сделки со следующей организацией. Заказы будут поступать неравномерно: в этом месяце к вам обратятся три компании, а в следующем — ни одной.

Поиск клиентов часто является самой сложной частью в работе независимого консультанта. Он требует преданности делу и большой сети контактов. Как правило, большинство приходит по рекомендации бывших коллег или клиентов. Чем больше людей знают консультанта и могут поручиться за его работу, тем больше клиентов у него будет. Таким образом, для того чтобы стать успешным консультантом, о его компетентности должны знать многие (в идеале те, у кого есть основания нанимать его). Чем разнообразнее ваша сеть деловых контактов с разными компаниями и индустриями, тем выше вероятность того, что работа будет приходить постоянно. Чтобы такая сеть сформировалась, нужно, чтобы консультант раньше работал со многими разными компаниями: либо часто менял место работы в начале карьеры, либо был консультантом в более крупной фирме.

Если вы добьетесь успеха как независимый консультант, у вас будет шанс нанять больше сотрудников и развивать бизнес. Сначала в компании есть только вы, но она может вырасти до команды из 5 человек или стать организацией из 100 человек. Вы как генеральный директор и ее основатель сможете вести свой бизнес в нужном вам направлении и с нужной культурой. Часть денег, которые приносит работа всех остальных консультантов, может сделать вас богатым. Хотя это случается редко, но работа независимого консультанта может стать самым прибыльным из всех путей, описанных в этой главе.

16.3.1. Преимущества работы в качестве независимого консультанта

Вы сами себе босс, а это значит, что вы можете решать, стоит ли браться за определенный проект, какой подход использовать и как представлять результаты. Вы

ни от кого не зависите; некоторым это дает ощущение свободы. Если вы сможете сократить расходы, поддерживать постоянный набор клиентов и вас ценят как специалиста, то в перспективе компания может стать довольно прибыльной. У вас есть возможность зарабатывать вдвое больше, чем в офисе, и это не предел. Если вы хотите поработать дома или взять выходной, у вас есть такая возможность, и никто не скажет вам слова против.

То, что вы делаете, принадлежит вам. Если вы придумали интересный метод решения задачи, то можете запатентовать его или продвигать как продукт своей компании. Никто не имеет права забрать вашу работу, тогда как если вы работаете на стороннюю фирму, она может заявить о ваших идеях как о своей интеллектуальной собственности. Если удастся создать портфолио полезных продуктов, оно сможет поддерживать вас в течение многих лет.

Консультации могут быть интересными! В том, чтобы летать по стране, помогать людям, предлагая им свои идеи, и делать все это от имени собственной компании, есть нечто захватывающее. Наличие множества людей, желающих платить за ваше время, говорит о том, что вы все делаете правильно. Предложить решение, которое нравится клиенту, и знать, что вы сделали всю работу сами, — бесценно.

16.3.2. Недостатки работы в качестве независимого консультанта

Недостатки работы в качестве независимого консультанта огромны (они настолько значительны, что в этом разделе мы будем выделять их жирным шрифтом):

- **Независимый консалтинг — это ужасный стресс.** Получение зарплаты за конкретный месяц зависит от того, будут ли с вами сотрудничать, а это часто связано с факторами, не зависящими от вас (например, с бюджетом компании). И наоборот, вы можете обнаружить, что у вас больше работы, чем вы можете выполнить, и нужно понять, какой проект запустить сейчас. Часто приходится продавать консалтинговые проекты, прежде чем вы получите полный доступ к данным, который нужен для того, чтобы решить, можно ли реализовать проект в принципе. Если проект нереализуем, то вам придется думать, что с этим делать. Есть тысяча вариантов, когда работа в качестве независимого консультанта не даст вам уснуть по ночам.
- **Независимый консалтинг может разорить вас.** Если вы полностью перейдете в эту сферу и не сможете занять свою нишу, то быстро потеряете деньги. Даже если вы подпишете контракт, крупные компании часто не платят на протяжении 90, а то и 120 дней с момента завершения работы, так что вы можете получить оплату аж через полгода после старта проекта. Если такие качели в денежных поступлениях вам не подходят, то вы не сможете быть консультантом.

- **Вам будет не к кому обратиться за помощью.** К вам обращаются с определенной задачей, которую человек не может решить самостоятельно, и если вы работаете в одиночку, то у вас не будет тех, с кем можно было бы посоветоваться и обменяться идеями. Вы сами по себе. Если у вас возникнут трудности с проведением анализа или с развертыванием модели, придется искать решение самостоятельно; в противном случае надо будет сказать клиенту, что у вас ничего не получилось.
- **Работа будет мало связана с Data Science.** Количество времени, которое вы потратите на маркетинг, переговоры, составление контрактов и ведение бухгалтерии, будет огромным по сравнению с часами непосредственной работы с данными. Недостаточно быть сильным дата-сайентистом; вся остальная работа необходима консалтинговой компании, чтобы выжить.

Вы решили уйти из Data Science

Последний из вариантов — вообще отказаться от Data Science. Может быть, вы поймете, что такая работа вам больше не интересна. Возможно, нагрузка не соответствует вашим потребностям в балансе между работой и личной жизнью. Или же вы оказались в ситуации, когда вам приходится применять Data Science в неэтичных целях и вы больше не можете заставлять себя это делать. Есть много причин, по которым эта сфера подходит не для всех, и в этом нет ничего плохого.

Трудно давать советы об уходе из определенной сферы, потому что все зависит от вашего дальнейшего выбора. Опыт работы в Data Science — это простой способ попасть в смежные области, такие как разработка ПО или инженерия. Уйти в другую сферу может быть сложнее. Как мы говорили в главе 6, вы можете рассказать о своих предыдущих обязанностях так, чтобы сделать их максимально близкими к интересующей области. То же правило работает и в обратную сторону: сделайте свою роль в Data Science похожей на роль в другой области, насколько это возможно.

Если вы все же уйдете, возможно, вы захотите вернуться позже. Если после ухода вы будете иногда брать небольшие проекты или стараться быть в курсе современных тенденций, то вам будет проще наверстать упущенное, когда вы будете готовы вернуться. В этом случае ситуация будет похожа на описанную в начале книги, только ваш опыт будет обширнее. Хотя вокруг Data Science сейчас очень много шума и о ней говорят как об очень востребованной сфере, не позволяйте этому заставлять вас чувствовать, будто вы должны остаться; главное, чтобы вы были счастливы. Делайте то, что вам нужно.

16.3.3. Как стать независимым консультантом

Для этой роли у вас должны быть сильные навыки в Data Science и опыт решения проблем без посторонней помощи. Вам также понадобятся специалисты, с которыми вы работали в нескольких компаниях или (что еще лучше) в крупной консалтинговой фирме, которые знают ваши способности и компетентность.

Вы можете прощупать почву, занимаясь фрилансом в свободное время. Создайте веб-сайт, разместите объявление на LinkedIn и сообщите, что готовы помогать людям. Если вам удастся найти клиентов, то вы узнаете о консалтинге больше благодаря фрилансу. Если вы поймете, что вам не хватает сил на работу по вечерам, то, скорее всего, вам не понравится быть консультантом. Если у вас не получится найти клиентов в качестве фрилансера, это говорит о недостаточно большой сети контактов: займитесь ее расширением в первую очередь.

Если вы поймете, что не справляетесь с основными обязанностями из-за большого объема внештатной работы, значит, сейчас отличное время, чтобы стать независимым консультантом. На этом этапе вы можете всецело сосредоточиться на консультировании, и, если у вас получится найти стартовый набор клиентов, можете уволиться с основной работы.

16.4. Выбор своего пути

В этой главе мы предложили три варианта развития карьеры в Data Science, но есть много других. Все эти пути могут быть довольно тернистыми, и по большей части вы не можете пойти по какому-либо из них, прежде чем возьмете на себя обязательство идти до конца. Как узнать, что подходит именно вам?

Правда в том, что вы не можете этого знать. Вы не можете знать, какой выбор «правильный», потому что правильного выбора попросту нет. Эти решения зависят от компаний, с которыми вы работаете, людей, которые вас окружают, и ваших личных интересов в конкретный момент жизни. Вы можете сделать только тот выбор, который вам нравится больше всего, и не слишком беспокоиться об упущенных возможностях.

Мы говорили об этом снова и снова. Точно так же как нет правильного способа освоить навыки работы с данными или какого-то одного правильного типа компании, куда можно было бы устроиться на работу, нет идеального способа ориентироваться в основных этапах вашей карьеры. Вы должны делать то, что лучше для вас, руководствуясь только имеющимися знаниями. Мы надеемся, что из этой главы вы узнали достаточно, чтобы у вас получилось легче выбрать для себя желаемую профессию.

16.5. Интервью с Анджелой Басса, руководителем отдела Data Science, инженерии данных и машинного обучения в iRobot

Анджела Басса (Angela Bassa) работает директором в iRobot, где она курирует инженерию данных, Data Science и МО во всей организации. Ранее она работала

консультантом, старшим руководителем и директором по аналитике. У нее есть степень бакалавра математики.

Какова повседневная жизнь руководителя?

Все зависит от того, насколько сложна структура организации, а это, в свою очередь, обычно зависит от ее размера. Если у вас в подчинении три человека, то они связаны между собой тремя каналами коммуникации, а если подчиненных семь, то каналов взаимодействия в разы больше. Если мне нужно координировать работу между разными продуктами, командами, целями и сроками, то потребуется много встреч и совещаний. Я трачу около трети дня на стратегическую координацию, чтобы убедиться, что мы работаем над правильными задачами, используем правильные методы и движемся к правильной цели. Еще треть тратится на взаимодействие с командой: как правило, я выступаю в роли наставника и помогаю разобраться в ситуации или даю обратную связь. Последняя треть — административные задачи. Например, все ли в порядке с бюджетом? Всем ли хватает денег на обучение и развитие, которое выбрал каждый сотрудник? Если планируется действительно интересная женская конференция и у меня в команде есть несколько кандидатов, хочу ли я спонсировать ее?

Какие признаки свидетельствуют о том, что пора перестать быть исполнителем?

Решение стать руководителем требует самоанализа, осознанности и широких взглядов. Момент, когда переход на должность руководителя с наибольшей вероятностью принесет вам успех, во многом определяется тем, что человек должен быть на своем месте (как в профессиональном, так и в личном плане). Руководитель — это другой род деятельности: у него другой набор навыков и профиль риска. Если вы ошиблись как исполнитель, то несете ответственность только за себя. Когда вы управляете другими людьми, ваша ответственность — это чей-то доступ к медицинскому страхованию или чья-то способность платить аренду. Но я действительно считаю, что любой может стать руководителем, и если вы к этому стремитесь, то вы наверняка отличный кандидат.

Придется ли когда-нибудь отказаться от роли исполнителя?

Data Science настолько новая сфера, что в ней до сих пор приходится делать непростой выбор среди тех, кто пришел в эту профессию. Многие из нас амбициозные и целеустремленные: мы выстраиваем новую карьеру, потому что принадлежим к тем людям, которые точно так же строят свою жизнь. Но если вы посмотрите на другие сферы, например бухгалтерское дело, там все иначе; можно очень долго работать на должности главного бухгалтера. Единственное, что может случиться, — вы достигнете потолка зарплаты или профессионального развития. Если это соответствует вашим карьерным целям, то не вижу ничего плохого в том, чтобы человек, которого все утраивает, оставался там, где он есть. Тем не менее

есть много талантливых профессионалов, которые стремятся построить карьеру в Data Science, поэтому если мне придется решать, кого продвигать, скорее всего, я выберу именно такого человека.

Что вы посоветуете тем, кто хочет быть техлидом, но не совсем готов к этому?

Найдите человека, с которым вы сможете открыто обсудить этот вопрос; одна голова хорошо, а две — лучше. Так вы получите четкую обратную связь и поймете, что вам необходимо прокачать. Человек, который уже стал успешным техлидом, может помочь вам понять, как к этому прийти. Исследуя свой собственный набор навыков, вы сможете заметить и пробелы в них. Парадокс: мы согласны с тем, что для роста требуется много общения и сотрудничества и что нужна целая деревня, чтобы вырастить ребенка, но в профессиональном плане мы ожидаем, что каждый должен быть в состоянии решать задачи самостоятельно. Лучший способ расти — это находить людей, которые поддерживают вас и смогут дать оценку вашей работе.

Ваш последний совет начинающим и младшим дата-сайентистам?

Мой первый совет — будьте скромнее. Легко возомнить себя царем горы. Вы можете думать, что Data Science — самая крутая профессия и любой работодатель должен сыпать лепестки роз к нашим ногам. Очень важно помнить, что для успеха продукта над ним работает много людей, и хотя наука о данных находится в центре внимания, это не означает, что она какая-то особенная или лучше, чем другие сферы деятельности.

Второй совет — будьте добрее к себе. Часто мы чересчур строги к себе, особенно потому, что сфера Data Science очень обширна и включает много разных направлений. Если вы хорошо разбираетесь в анализе, но хуже знаете машинное обучение, то можете беспокоиться, что вы не «настоящий» специалист по данным. Но это не так! Есть множество способов проявить себя.

Итоги

- Управление отлично подходит тем, кто хочет помогать другим и готов отказаться от Data Science. Этот путь может привести к высокому положению в компании.
- Ведущий дата-сайентитст должен руководить технической стороной и отвечать за других. Эта роль — отличный вариант, чтобы продолжать решать технические вопросы и помогать другим.
- Независимый консалтинг — это очень напряженный и рискованный путь, но в перспективе он может принести свои плоды. Чтобы получать заказы постоянно, вам понадобится сильная сеть контактов.

Материалы к главам 13–16

Книги

The Design of Everyday Things, Don Norman (Basic Books)

Эта философская книга содержит идеи из области дизайна и рассказывает о том, как думать о дизайне в любом виде деятельности. Умение понимать пользователя и то, как дизайн влияет на его действия, имеет решающее значение для успеха продукта. Прочитав эту книгу, вы научитесь лучше понимать потребности стейкхолдеров и снизите вероятность провала вашего проекта, потому что это не то, чего хочет заказчик.

Self-Compassion: The Proven Power of Being Kind to Yourself, Kristin Neff, PhD (HarperCollins Publishers)

Если часть главы 13, где рассказывалось, как люди истязают себя, если их проект провалился, не оставила вас равнодушным, прочитайте эту книгу. Она подробно описывает борьбу автора с самокритикой и ее путь любви к себе. Этот путь должны пройти многие дата-сайентисты, и книга станет отличным проводником.

Demystifying Public Speaking, Lara Hogan (A Book Apart)

Лара Хоган, известный спикер в области инженерного управления, написала эту книгу, чтобы дать практические советы, которые помогут людям начать говорить. В этой короткой и увлекательной книге она описывает тактики на все случаи жизни: от выбора темы до выступления и борьбы с волнением.

R Packages, 2nd ed., Jennifer Bryan and Hadley Wickham (O'Reilly Media)

В этой книге подробно рассказывается, как создать пакет R, усовершенствовать рабочий процесс и как работать на благо сообщества. В январе 2020 года эта книга все еще находилась в процессе написания, но вы можете найти незавершенную копию по ссылке https://r-pkgs.org.

Resilient Management, Lara Hogan (A Book Apart)

Эта книга — отличное краткое руководство для любого руководителя, который недавно приступил к новой должности. В ней вы найдете советы по знакомству с членами команды, узнаете о наставничестве и о том, как определять ожидания и решать сложные задачи. Даже если вы не планируете управлять людьми, эта книга будет очень полезна, если вы только начинаете руководить проектами или испытываете трудности в общении с командой.

The Manager's Path: A Guide for Tech Leaders Navigating Growth and Change, Camille Fournier (O'Reilly Media)

Если вы хотите, чтобы вас повысили до технического руководителя или менеджера, эта книга идеально подойдет. Камиль Фурнье, бывший технический директор Rent the Runway, рассказывает о том, как оценивать работу с точки зрения руководителя, а не исполнителя. Эта книга поможет вам понять идеи, на осмысление которых могут потребоваться годы.

The E-Myth Revisited, Michael E. Gerber (HarperCollins Publishers)

Хотя эта книга совершенно не связана с Data Science, она является отличным ресурсом для людей, которые рассматривают возможность работы независимыми консультантами или хотят начать собственное дело. В ней рассказывается, как должны меняться представления о работе по мере того, как вы ведете бизнес. Вместо того чтобы сосредоточиться на выполнении задач, вы должны стремиться систематизировать все процессы, чтобы бизнес продолжал функционировать непрерывно.

High Output Management, Andrew S. Grove (Vintage)

Вести бизнес хорошо — это суровое испытание, требующее развитого стратегического мышления. В этой книге рассматриваются простые примеры вроде доставки завтрака; она станет полезным ресурсом для людей, которые хотят больше узнать об управлении организацией. Книга была написана в 1983 году, но с тех пор в нее вносили изменения, и возраст на ней совершенно не отразился.

Блоги

«Making peace with personal branding», Rachel Thomas

https://www.fast.ai/2017/12/18/personal-brand

Рэйчел Томас отлично справляется с продвижением себя и в Data Science, и в соцсетях; в этом посте рассказывается, как это делать, сохраняя чувство комфорта.

Блог Лары Хоган

https://larahogan.me/blog

Мы советовали прочитать две книги Лары Хоган; ее блог не менее интересен и полон отличных советов о мягких навыках, необходимых для достижения успеха. Многие из ее постов сосредоточены вокруг руководителей, но она также дает универсальные советы, которые подходят всем: в том числе что делать, если руководитель не поддерживает вас, как дать обратную связь и как справиться со своими эмоциями, когда вы думаете: «Почему руководитель не может просто…?»

«The art of slide design», Melinda Seckington

https://missgeeky.com/2017/08/04/the-art-of-slide-design

Эта серия из пяти постов (ссылка на каждый последующий пост дается внизу) представляет собой мастер-класс по созданию эффективных слайдов. Секингтон называет принципы дизайна слайдов — максимизировать сигнал, минимизировать шум, выделять важную информацию, показывать и говорить и быть системным — и иллюстрирует их множеством примеров и контрпримеров.

«Overcoming social anxiety to attend user groups», Steph Locke

https://itsalocke.com/blog/overcoming-social-anxiety-to-attend-user-groups

Если социофобия мешает вам посещать встречи или конференции, почитайте этот пост. Она затрагивает такие распространенные проблемы, как «Я никого не знаю» или «Как мне разговаривать с людьми?» и дает краткие практические советы.

«How to ask for a promotion», Rebecca Knight

https://hbr.org/2018/01/how-to-ask-for-a-promotion

В этой статье приводятся советы двух тренеров по лидерству о том, как просить о повышении. Каждый совет, такой как «Посадите зернышко» и «Проведите небольшое исследование», содержит абзац с конкретными примерами.

Эпилог

Что ж, мы затронули много тем. Мы начали с определения Data Science и того, какие навыки вам необходимы; рассказали, как подготовиться к интервью и получить работу; обсудили варианты карьерного роста в этой области. Все шестнадцать глав этой книги охватывают разные темы, с понимания различных типов компаний до создания модульного тестирования для продукционной модели и вопросов о том, как стать руководителем.

Если посмотреть на книгу в целом, то можно заметить, что в ней прослеживаются некоторые тенденции. Эти уроки применимы ко всем этапам пути дата-сайентиста. Что касается нас, то мы взяли за основу три следующих правила, следуя которым нам удалось продолжить карьерный рост:

- *Дата-сайентист должен уметь общаться*. Не раз люди, у которых мы брали интервью для книги, упоминали, что достигли успеха именно благодаря эффективному общению. Будь то составление отчета для руководителя, взаимодействие с командой инженеров по модели или умение говорить так, чтобы вас понимали нетехнические специалисты, — общение может помочь вам в поиске работы и взаимодействии с другими людьми на новом месте.
- *Дата-сайентист должен проявлять инициативу*. Чрезвычайно редко специалист по данным получает четко сформулированную задачу и инструменты для ее решения. Вместо этого ему необходимо самостоятельно пытаться найти данные, придумывать новые идеи для моделей и проводить эксперименты. Проявление инициативы и создание портфолио также поможет вам получить работу. Чем больше задач вы решите и чем активнее будете проявлять инициативу, тем лучше.
- *Дата-сайентисту нужно сообщество*. Никто не делает карьеру без посторонней помощи, но как представителям новой и быстрорастущей области специалистам по данным нужны прочные профессиональные связи, которые могут принимать разные формы. Спонсор может порекомендовать вас в качестве спикера на митапе, благодаря чему через два года вы уже будете выступать на международной конференции. Наставник может оценить ваше резюме

и порекомендовать вас на открытую вакансию в компании. Менеджер может помочь преодолеть разрыв со стейкхолдерами и предложить варианты для личностного роста. Или коллега может просто посочувствовать и поднять настроение, если у вас был тяжелый день. Стоит потратить время на выстраивание взаимоотношений: они помогут вам решить множество проблем, с которыми вы столкнетесь на карьерном пути.

Мы надеемся, что вам понравилось читать эту книгу; нам определенно понравилось писать ее. При написании мы обнаружили, что во многих случаях описывали собственный опыт. Несколько раз в процессе мы даже перечитывали некоторые места, чтобы лучше понять принятые нами карьерные решения. Мы желаем вам удачи на пути к карьере в Data Science!

Приложение. Вопросы интервью

Часто самое полезное при подготовке к собеседованию — это понять, на что оно будет похоже. От умения быстро реагировать и отвечать на вопросы может зависеть получение работы. Далее мы предлагаем примеры вопросов, которые могут задать на интервью, чтобы вы могли их понять и обдумать. Советуем рассматривать вопросы в сочетании с главой 7, где описывается процесс собеседования в целом.

Вопросы в приложении делятся на пять категорий:

- написание кода и разработка ПО;
- SQL и базы данных;
- статистика и машинное обучение;
- поведенческие вопросы;
- вопросы на логику.

Data Science охватывает широкий спектр тем, поэтому невозможно изучить тысячи потенциальный вопросов, чтобы подготовиться к интервью. В одной компании вас попросят инвертировать двоичное дерево, а в другой зададут только поведенческие вопросы и спросят о Python. Именно поэтому мы рекомендуем выяснить заранее, к каким вопросам следует подготовиться. Конечно, вы не получите полный перечень, но менеджер, набирающий команду, или рекрутер должны дать вам общее представление о том, чему следует уделить особое внимание при подготовке. Например, вам могут сказать: «На первом интервью нужно ответить на несколько вопросов по SQL на вайтборде. Затем вам предстоит пройти два поведенческих интервью подряд: одно с инженером, а другое с дата-сайентистом. В конце один из наших инженеров МО спросит о ваших предыдущих проектах в Data Science».

Крайне маловероятно, что в процессе поиска работы вам встретятся только вопросы из этого приложения. Поэтому мы не только предоставили варианты ответов на каждый вопрос (включая непосредственно текст ответа и пример кода), но и отметили, что, на наш взгляд, делает их оптимальными. Все ответы даны от первого лица гипотетического специалиста по данным, опыт работы которого

аналогичен опыту авторов этой книги. В некоторых вопросах мы ссылаемся на опыт, полученный на предыдущих местах работы; постарайтесь придумать собственные примеры для таких вопросов.

Некоторые вопросы связаны с нашим личным опытом, приобретенным в ходе множества собеседований, а другие были предоставлены нашими коллегами. Большое спасибо всем, кто внес свой вклад и помог сделать это приложение полезнее!

A.1. Написание кода и разработка ПО

A.1.1. FizzBuzz

Напишите программу, которая выводит на экран числа от 1 до 100. При этом вместо чисел, кратных 3, программа должна выводить слово «Fizz», а вместо чисел, кратных 5, — слово «Buzz». Если число кратно 15, то программа должна выводить слово «FizzBuzz».

ПРИМЕР ОТВЕТА

Ниже представлен псевдокод одного из вариантов решения:

```
for (i in 1 to 100) {
    if (i mod 15) {
        print("FizzBuzz")
    } else if (i mod 5) {
        print("Buzz")
    } else if (i mod 3) {
        print("Fizz")
    } else {
        print(i)
    }
}
```

Программа выполняет итерацию чисел от 1 до 100. Для каждой итерации она сначала проверяет, делится ли число на 15, и если это так, то выводится слово «FizzBuzz». Если нет, она проверяет, делится ли число на 5: если да, то выводится «Buzz». Если число не делится на 5, то программа проверяет, делится ли оно на 3, и выводит на экран «Fizz», если это так. Если число не делится ни на один из вариантов, программа выводит на экран число.

ПРИМЕЧАНИЯ

Это чрезвычайно распространенная задача, которую задают в интервью в сфере разработки ПО. Она была придумана Имраном Гори и популяризирована Джеф-

фом Этвудом (https://blog.codinghorror.com/why-cant-programmers-program), так что именно этот вопрос можно часто услышать в интервью на вакансию в Data Science. Две основные задачи, которые в нем решаются, — выяснить, как итерировать набор всех чисел (в примере мы использовали цикл `for`) и как проверять, что должно быть выведено для каждого числа. Распространенная ошибка при решении этой задачи — проверять, делится ли число на 3 или 5, прежде чем узнать, делится ли оно на 15; любое число, которое делится на 15, также делится на 3 или 5. То есть если сначала проверяется кратность 3 или 5, то программа может вывести «Fizz» или «Buzz» там, где должно быть выведено «FizzBuzz».

Мы предложили очень простое решение, но его можно усовершенствовать. В некоторых языках, включая R и Python, можно применить подход более чистого функционального программирования, используя purrr в R или списковое включение в Python. Еще можно создать обобщенную функцию, которая в качестве входных данных принимает список кратных значений для проверки и слова для вывода с этими кратными числами и выводит любой список. В зависимости от того, как проходит интервью, вы можете обсудить варианты разных решений.

Если интересно, гляньте версию FizzBuzz Enterprise Edition (https://github.com/EnterpriseQualityCoding/FizzBuzzEnterpriseEdition) или модель машинного обучения FizzBuzz TensorFlow (https://joelgrus.com/2016/05/23/fizz-buzz-in-tensorflow).

A.1.2. Скажите, является ли число простым

Напишите функцию, которая возвращает истину, если переданное число простое, и ложь, если не простое. Предположим, что у вас нет встроенной функции, чтобы проверить, является ли число простым.

ПРИМЕР ОТВЕТА

Ниже представлен псевдокод варианта решения:

```
is_prime = function(n) {
    for (i in 2 to n / 2) {
        if ((n mod i) == 0) {
            return FALSE
        }
    }
    return TRUE
}
```

Простое число — это такое, которое делится только на 1 и на само себя. Программа итерирует все числа от 2 до половины заданного числа и проверяет, делится ли на них заданное число. Если это так, то функция возвращает `false` и останавливается. Если цикл `for` проходит без остановки, функция возвращает `true`.

ПРИМЕЧАНИЯ

Подобно FizzBuzz, эта задача проверяет, сможете ли вы написать цикл `for` и функцию. Вы также должны знать, как остановить итерацию, когда условие выполнено, чтобы при этом можно было безопасно вернуть `true` в конце, если цикл `for` завершится. Помните о небольших хитростях, например вы должны понимать, что не нужно проверять, делится ли число на все меньшие числа или на его половину; проверяйте только те, которые меньше, чем его квадратный корень. Смысл задачи в том, чтобы проверить, можете ли вы написать работающую функцию.

A.1.3. Работа с Git

Расскажите об опыте использования Git для совместной работы над проектом.
Алекс Хейс

ПРИМЕР ОТВЕТА

На последнем месте работы я создал пакет R, funneljoin на спринте совместно с двумя коллегами, при этом с самого начала мы использовали Git. Первый час мы применяли метод парного программирования на одном компьютере, а затем составили список задач, которые нужно было разделить между нами. Каждый из нас создал отдельную ветку для работы над задачами, что позволило потом легко все объединить. Использование Git исключает возможность случайно переписать чужую работу. Выполняя коммиты часто и на ранних этапах, мы знали, что всегда можем вернуться к предыдущему способу реализации функции, если решим, что он лучше. Наконец, использование GitHub означало, что впоследствии любой сотрудник компании сможет загрузить пакет и сразу же начать его использовать.

С этого дня я занимался администрированием пакета. Я продолжаю использовать элементы Git (например, ветки), чтобы создавать прототипы функций, не объединяя их, пока они не будут тщательно протестированы.

ПРИМЕЧАНИЯ

Если вам задали этот вопрос, но вы раньше не использовали Git для совместной работы, расскажите, как вы применяли его в собственном проекте. Этот вопрос задают для того, чтобы понять, есть ли у вас опыт работы с Git, а также можете ли вы объяснить, какие функции использовали (например, ветки или разветвление), пусть и в личных целях. Расскажите, как бы вы доработали свои методы для совместного использования, например скажите, что можно добавить больше веток или сохранять структурированную систему коммита.

Если вы раньше никогда не использовали Git, скажите об этом честно. Но мы настоятельно рекомендуем ознакомиться с этой темой, прежде чем отправляться на несколько собеседований.

A.1.4. Технологические решения

Как выбрать технический стек, если вы начинаете работу с нуля?

Хизер Нолис

ПРИМЕР ОТВЕТА

Интересный вопрос: все зависит непосредственно от реализуемого проекта. Я пытаюсь найти баланс между простотой решения задачи лично для меня и тем, как другие потом будут работать с моим решением. Позвольте мне привести два примера выбора технического стека, а также рассказать о том, чему меня это научило.

В начале карьеры мне пришлось работать над одним из проектов, для которого требовалось разработать новый продукт с нуля, и я был единственным датасайентистом. Я решил работать со стеком .NET и языком программирования F#, так как мне были хорошо знакомы данные технологии и они позволяли быстро запустить продукт в работу. Существенным недостатком этого решения стало то, что F# был редко используемым языком, и когда перед нами встал вопрос о том, чтобы нанять еще одного специалиста по данным, оказалось, что это не так просто, поскольку мы не могли найти человека, у которого были необходимые знания. Сейчас, вспоминая этот случай, я понимаю, что выбор в пользу .NET и F# был не самым правильным решением.

Что касается недавних проектов, мне поручили создать машинное обучение. Я работал в команде с инженерами, занимающимися микросервисами, поэтому решил создать REST API на R в качестве Docker-контейнера. Несмотря на то что у меня не было опыта в работе с Docker-контейнерами, я знал, что они наиболее простые в обслуживании. В процессе работы мне удалось многое узнать о Docker контейнерах, а результат моей работы оказался хорошо интегрируемым.

ПРИМЕЧАНИЯ

Когда Хизер Нолис проводит интервью с кандидатами в T-Mobile, она обычно предлагает им выбрать проект и описать принятые по нему решения, а также рассказать, что бы кандидат сделал по-другому, уже зная все тонкости. Не факт, что на интервью вам зададут тот же вопрос, но если вы примените такой прием, это отметят.

Независимо от того, как вы будете отвечать, нужно подробно описать решения, которые вам приходилось принимать в процессе работы (можете также рассказать

о сторонних проектах и курсовых работах). Цель этого вопроса — понять, насколько хорошо вы подошли к вопросу выбора подходящих технологий. Не страшно, если вы выбрали не тот стек, но сделали правильные выводы. На самом деле умение делать выводы даже лучше умения делать все правильно с первого раза, так как это доказывает, что вы способны меняться.

A.1.5. Часто используемый пакет/библиотека

Какой пакет R или библиотеку Python вы использовали часто и почему?

ПРИМЕР ОТВЕТА

Я использовал набор tidyverse в R, и, хотя это не один пакет, он мне нравится. С помощью пакетов можно пройти весь путь от чтения данных до их очистки и преобразования, визуализации и моделирования.

Мне особенно нравится dplyr, поскольку благодаря его объединенному пакету я могу писать тот же код независимо от того, работаю ли я над локальной или удаленной таблицей, так как он переводит код dbplyr в SQL. На последнем месте работы я использовал dbplyr, благодаря чему можно было оставаться в RStudio на протяжении всего рабочего процесса, даже несмотря на то, что все наши данные хранились в Amazon Redshift и для доступа к ним требовались запросы SQL. Я использовал этот пакет для составления сводных таблиц и фильтрации, а затем извлекал данные локально, если мне нужно было выполнять более сложные операции или визуализации.

В целом мне действительно нравится философия Хэдли Уикхэма (Hadley Wickham), основного разработчика tidyverse: в программировании больше всего времени уходит на обдумывание, а не вычисления, и вы должны создавать сочетаемые инструменты, работающие без сбоев, и уметь быстро облекать свои мысли в код.

ПРИМЕЧАНИЯ

Вряд ли интервьюер ждет здесь конкретного ответа. Скорее он хочет понять, (а) достаточно ли у вас опыта в программировании на любом из языков, где есть часто используемый пакет, и (б) можете ли вы объяснить, как и почему вы используете этот пакет. Этот ответ также дает интервьюеру представление о том, какой работой вы занимаетесь ежедневно. Не забудьте объяснить, что делает пакет, особенно если он узконаправленный. Если для решения этой задачи есть другой, более распространенный пакет, возможно, стоит объяснить свой выбор, поскольку это показывает, что вы шире понимаете альтернативные варианты.

Не старайтесь выбрать самую «продвинутую» библиотеку, чтобы произвести впечатление на интервьюера. В идеале это один из самых простых и (потенциально) самых интересных вопросов, поэтому не зацикливайтесь на нем.

A.1.6. R Markdown или Jupyter Notebooks

Что такое файл R Markdown или Jupyter Notebook? Зачем использовать файл R Markdown или Jupyter Notebook вместо сценария R или Python? Когда лучше использовать сценарий?

ПРИМЕР ОТВЕТА

Я отвечу на этот вопрос на примере R и R Markdown, но для Python и Jupyter Notebooks основная идея та же. Файлы R Markdown — это способы написания кода R, которые позволяют размещать текст и форматировать код. В некотором смысле они объединяют код и результаты анализа с нарративом и идеями анализа. Используя файл R Markdown, вы получаете анализ, который воспроизводится легче, чем необработанный код R, плюс к нему будет идти отдельная документация по анализу. В идеале R Markdown должен быть отформатирован настолько чисто, чтобы при рендеринге выходного файла вы могли передать стейкхолдеру полученный результат в виде HTML, документа Word или PDF.

Файлы R Markdown отлично подходят для воспроизводимого анализа, но они не очень пригодятся, если код пишется для развертывания или использования в других местах. Допустим, у вас есть список функций, которые вы хотите использовать в нескольких других местах (например, для загрузки данных из файла). Возможно, есть смысл написать сценарий R, который создает все функции, и сохранить его отдельно от анализа. Или если вы хотите использовать R с пакетом plumber для создания веб-API, для этого не нужен файл R Markdown.

ПРИМЕЧАНИЯ

Задавая этот вопрос, интервьюер проверяет, есть ли у вас опыт проведения воспроизводимого анализа. Многие специалисты, использующие R или Python, пишут сценарии бессистемно и не думают о том, как потом будут делиться результатами с другими. Показывая, что вы понимаете суть R Markdown и Jupyter Notebooks, вы доказываете, что думаете о том, как сделать код удобным в использовании. Если вы не работали с файлами R Markdown или Jupyter Notebooks, обязательно попробуйте сделать это.

Не думайте, будто нужно обязательно понимать обе версии: и R, и Python, подойдет любая.

A.1.7. Когда следует писать функции или пакеты/библиотеки?

В какой момент нужно преобразовать код в функцию? Когда нужно преобразовать его в пакет или библиотеку?

ПРИМЕР ОТВЕТА

Обычно, как только я замечаю, что копирую и вставляю код, это говорит о том, что следует превратить его в функцию. Например, если мне нужно запустить код для трех разных датасетов, то нужно создать функцию и применить ее к каждому из них, а не копировать код трижды. Библиотека purrr в R или списковое включение в Python позволяет легко применить функцию много раз.

Я заметил, что пакеты и библиотеки лучше всего использовать в случаях, если у вас есть код, охватывающий несколько разных проектов команды. На текущем месте работы мы храним много данных в S3, но хотим анализировать их локально. Вместо того чтобы копировать и вставлять функции для доступа к коду каждого проекта, я создал библиотеку, которую можно было вызывать из любого места. Недостатком библиотек является то, что если вы их измените, то придется менять все проекты, в которых они используются, но для основных функций этот подход часто бывает целесообразным.

ПРИМЕЧАНИЯ

Отчасти это простой вопрос, потому что на него есть правильный ответ: «Как можно чаще». Как правило, многократное копирование и вставка кода — неудачный метод; дата-сайентист должен создавать функции так, чтобы код было легче читать и понимать. Поэтому покажите, что вы понимаете важность повторного использования кода в качестве функций или пакетов. Вам когда-нибудь приходилось создавать функцию и использовать ее много раз? А библиотеку? Приводите как можно больше примеров.

A.1.8. Пример работы с данными в R/Python

Перед вами таблица с твитами. В данных есть учетная запись, с которой был отправлен твит, текст, количество лайков и дата отправки. Напишите сценарий, чтобы получить таблицу, в которой будет выделена строка для каждого человека, столбец min_likes с минимальным количеством лайков, которые он получил, и столбец nb_tweets с общим количеством твитов. Таблица должна включать только твиты, отправленные после 1 сентября 2019 года. А еще сначала необходимо удалить все дубликаты в таблице.

account_name	text	nb-likes	date
@vboykis	Data Science — это…	50	2019-10-01
@Randy_Au	Сложно, когда…	23	2019-05-01
@rchang	Немного новостей…	35	2019-01-01
@vboykis	Моя рассылка…	42	2019-11-23
@drob	Мой лучший совет…	62	2019-11-01
…	…	…	…

ПРИМЕР ОТВЕТА НА R

```
tweets %>%
  filter(date > “2019-09-01”) %>%
  distinct() %>%
  group_by(account_name) %>%
  summarize(nb_tweets = n(), min_likes = min(nb_likes))
```

ПРИМЕР ОТВЕТА НА PYTHON

```
tweets = tweets[tweets.date > "2019-09-01"].
drop_duplicates().
groupby("account_name")

tweets['nb_likes'].agg(nb_tweets="count", min_likes="min")
```

ПРИМЕЧАНИЯ

Этот тип вопросов представляет собой некую смесь вопросов по FizzBuzz и простым числам (проверяет умение работать с R/Python) и запросов SQL (анализ данных). Этот вопрос должен быть относительно простым для тех, кто прежде имел дело с анализом данных. Но вам может встретиться и более сложное задание (например, преобразовать символьный столбец в столбец с датой или изменить формат данных с длинного на широкий), и вы можете попросту забыть, как это делается. Если вы не помните, как решать конкретную задачу, просто скажите: «Я не помню точного синтаксиса для X, поэтому я вставлю туда какой-нибудь временно замещающий псевдокод» и двигайтесь дальше. Не зацикливайтесь слишком долго на одной части. Если вопрос действительно подразумевает что-то нестандартное, скорее всего, интервьюер сочтет эту часть дополнительной, а не обязательной для перехода к следующему этапу.

A.2. SQL и базы данных

A.2.1. Типы соединений

Объясните разницу между левым и внутренним соединениями.
Лудамила Джанда (Ludamila Janda) и Аянти Дж. (Ayanthi G.)

ПРИМЕР ОТВЕТА

Соединения — это способы объединения данных из двух разных таблиц (левой и правой) в новую. Принцип работы соединений заключается в объединении строк между двумя таблицами; набор ключевых столбцов используется для поиска данных в двух одинаковых таблицах, которые необходимо соединить. В случае левого соединения каждая строка из левой таблицы появляется в полученной таблице, но строки из правой появляются, только если значения в их ключевых столбцах отображаются в левой. Но при внутреннем соединении обе строки из левой и правой таблиц появляются, только если есть совпадающая строка в другой таблице.

На практике вы можете воспринимать левое соединение как подсоединение данных из правой таблицы к левой, если она существует (например, с использованием правой таблицы в качестве подстановочной). Внутреннее соединение больше похоже на поиск всех общих данных и создание новой таблицы только из пар.

ПРИМЕЧАНИЯ

Лудамила Джанда любит задавать этот вопрос при наборе кандидатов на роли джуниоров, потому что это не вопрос с подвохом, а необходимые знания. Она считает, что можно многое узнать по тому, как кандидат решает дать ответ. Правильно ответить можно по-разному: процитировать учебник, используя сложные выражения, или объяснить все простым языком, не вдаваясь в крайности.

Обратите внимание, что в нашем варианте ответа мы не упоминали какие-либо сложности, связанные с появлением повторяющихся строк в данных. Возможно, о них стоит упомянуть, потому что повторяющиеся строки могут повлиять на результат, но с большей долей вероятности эта информация может увести вас от мысли, которую вы пытаетесь донести до интервьюера.

A.2.2. Загрузка данных в SQL

Назовите несколько различных способов загрузки данных в БД, которые следует использовать в первую очередь, и каковы преимущества и недостатки каждого из них?
Аянти Дж.

ПРИМЕР ОТВЕТА

Есть много способов загрузить данные в базу, и выбор зависит в первую очередь от того, где они находятся. Если они расположены в плоском файле вроде CSV, то во многих версиях SQL есть программы для их импорта. Например, в SQL Server 2017 есть «Мастер импорта и экспорта». Эти инструменты просты в использовании, но их нелегко кастомизировать или воспроизвести. Если данные поступают из другой среды, такой как R или Python, существуют драйверы, позволяющие передавать данные в SQL. Например, драйвер ODBC можно использовать вместе с пакетом DBI в R для перемещения данных из R в SQL. Эти методы более воспроизводимы и программируемы с точки зрения реализации, но они требуют передачи данных в R или Python.

ПРИМЕЧАНИЯ

Этот вопрос проверяет, приходилось ли вам раньше загружать данные в БД. Если вы это уже делали, описать процесс не составит труда. Если нет, интервьюер может посчитать, что у вас недостаточно опыта.

Часть вопроса о преимуществах и недостатках разных методов проверяет, понимаете ли вы, что в зависимости от ситуации определенные инструменты работают лучше. Иногда GUI — хорошее и простое решение, если у вас всего один файл. В других случаях лучше создать полностью автоматизированный сценарий для непрерывной загрузки. Чем детальнее вы сможете объяснить, что и когда использовать, тем лучше.

A.2.3. Пример SQL-запроса

Перед вами школьная таблица TABLE_A, содержащая оценки от 0 до 100, выставленные ученикам нескольких классов. Как бы вы рассчитали максимальную оценку в каждом классе?

Класс	Ученик	Оценка
Математика	Эмбер Нолис	100
Математика	Майк Берковиц	90
Литература	Аманда Листон	97
Испанский	Лаура Бетанку	93
Литература	Эбби Робинсон	93
...	...	...

ПРИМЕР ОТВЕТА

Перед вами запрос для поиска наивысшей оценки в каждом классе:

```
SELECT CLASS, MAX(GRADE)
INTO TABLE_B
FROM TABLE_A
GROUP BY CLASS
```

Этот запрос группирует данные по каждому классу, а затем находит максимальное значение. Он дополнительно сохраняет результаты в новую таблицу (TABLE_B), чтобы их можно было запросить позже.

ПРИМЕЧАНИЯ

Это один из простейших вопросов по SQL; он проверяет, имеете ли вы базовое представление о группировке в SQL. Соискатели обычно путаются в нем из-за непонимания, что именно нужно группировать (в данном случае переменную класса), или же считают вопрос настолько простым, что начинают слишком усложнять ответ и упускают простое решение. Если вы услышали вопрос, который кажется вам слишком легким, вполне возможно, что так оно и есть. Если это решение не кажется вам очевидным, то самое время повторить, как группирующие переменные работают в SQL. Наконец, строка INTO TABLE_B была совершенно необязательной, но она хорошо подготовит вас к следующему вопросу.

A.2.4. Пример SQL. Продолжение

Рассмотрим таблицу из предыдущего вопроса. Что, если нужно было найти не только самую высокую оценку в каждом классе, но и учащегося, получившего ее?

ПРИМЕР ОТВЕТА

Предполагая, что у нас есть результат предыдущего вопроса, сохраненный в TABLE_B, мы можем использовать его в этом решении:

```
SELECT a.CLASS, a.GRADE, a.STUDENT
FROM TABLE_A a
INNER JOIN TABLE_B b ON a.CLASS = b.CLASS AND a.GRADE = b.GRADE
```

Этот запрос выбирает всех учащихся и их оценки из исходной таблицы TABLE_A, где указаны классы и оценки, которые отображаются в таблице максимальных значений TABLE_B. Внутреннее соединение работает как фильтр, сохраняя только максимальные комбинации классов/оценок, потому что только в этом случае оценка появляется в таблице TABLE_B. В качестве альтернативы

можно было бы использовать подзапрос, чтобы сделать то же самое, не вызывая таблицу TABLE_B:

```
SELECT a.CLASS, a.GRADE, a.STUDENT
FROM TABLE_A a
INNER JOIN (
  SELECT CLASS, MAX(GRADE)
  FROM TABLE_A GROUP BY CLASS) b
ON a.CLASS = b.CLASS AND a.GRADE = b.GRADE
```

ПРИМЕЧАНИЯ

Хотя эту задачу можно решить несколькими способами, любое решение почти наверняка требует более одного запроса из таблицы TABLE_A, поэтому этот вопрос может легко поставить в тупик. На бумаге решение может показаться легким, но додуматься до него во время интервью может быть непросто. Даже если вы ответили неправильно, это не значит, что вы провалились.

Решение не дает каких-либо специфичных случаев для привязки к максимальному значению. В примере будут возвращены строки для несколько студентов. Возможно, стоит указать на этот факт интервьюеру, потому что так вы показываете, что обращаете внимание на пограничные случаи.

A.2.5. Типы данных

В чем заключается недостаток хранения столбца дат в виде строк в базе данных? Например, что, если в SQL мы сохраним столбец дат как `VARCHAR(MAX)`*, а не как* `DATE`*?*

ПРИМЕР ОТВЕТА

Хранение дат в виде строк вместо чисел (например, хранение даты 20 марта 2019 года в виде строки «`20.03.2019`») встречается в БД достаточно часто. Несмотря на то что такой формат позволяет не потерять какую-либо информацию, вы можете столкнуться с падением производительности. Во-первых, если данные не хранятся как тип `DATE`, нельзя использовать для них функцию `MONTH ()`. А еще не получится найти различия между двумя датами или минимальную дату в столбце.

Эта проблема часто возникает, если вы загружаете данные в БД или очищаете ее. Чем раньше вы сможете правильно отформатировать данные, тем проще будет проводить анализ. Можно исправить подобные ситуации, используя функции вроде `CAST`. При этом если вы загружаете данные с сотнями столбцов, где много таких, которыми вы никогда не воспользуетесь, возможно, не стоит тратить время на устранение всех проблем.

ПРИМЕЧАНИЯ

Вопрос о неправильном выборе типа хранения данных задают очень часто. Такое можно встретить не только в БД, но и в плоских файлах или в таблицах в средах R и Python. С помощью этого вопроса пытаются понять, считаете ли вы такие вещи нормой или нет, а еще хотят узнать, что вы будете делать, чтобы все исправить. Умение отвечать на подобные вопросы должно сформироваться естественным образом, если в ходе работы над проектом по анализу вам приходится чистить данные.

A.3. Статистика и машинное обучение

A.3.1. Статистические термины

Объясните восьмилетнему ребенку термины «среднее арифметическое», «медиана» и «мода».

Аллан Батлер

ПРИМЕР ОТВЕТА

Среднее арифметическое, *медиана* и *мода* — это три разных типа средних значений. Средние значения дают нам понять что-то о целом наборе чисел с помощью всего одного числа, которое дает какой-то вывод о всем наборе.

Предположим, мы провели опрос в классе, чтобы узнать, сколько братьев и сестер у каждого ученика. В классе пять человек, и допустим, вы обнаружили, что у одного ученика вообще нет братьев и сестер, у одного — только один брат или сестра, еще у одного — два, а у двух других — пять.

Мода — это значение, которое встречается наиболее часто. В данном случае это 5, поскольку сразу у двух учеников есть пять братьев и сестер по сравнению с одним, у которого любой другой вариант.

Чтобы получить *среднее арифметическое*, нужно взять общее количество братьев и сестер и разделить его на количество учеников. В этом случае складываем 0 + 1 * 1 + 1 * 2 + 5 * 2 = 13. В классе пять учеников, поэтому среднее арифметическое получаем так: 13/5 = 2,6.

Давайте теперь выстроим все числа в ряд, от самого маленького к самому большому. *Медиана* — это число, которое будет стоять в середине этого ряда. Запишем по порядку 0, 1, 2, 5, 5. В нашем случае в самой середине находится третье число, а это означает, что медиана равна 2.

Мы видим, что три типа средних значений представлены разными числами. Когда какое использовать? Среднее арифметическое значение распространено больше всего, но если есть выбросы, нужна медиана. Предположим, у одного человека было 1000 братьев и сестер! Внезапно среднее арифметическое значение

становится намного больше, но на самом деле у большинства людей нет столько братьев и сестер. Что касается медианы, она остается прежней.

ПРИМЕЧАНИЯ

Маловероятно, что кандидат на должность дата-сайентиста не будет знать разницы между типами средних значений. Этот вопрос на самом деле проверяет ваши коммуникативные навыки, а не то, насколько правильно вы понимаете определения (хотя, если вы ошибетесь, это звоночек для интервьюера). В данном случае мы привели простой пример, с которым в реальной жизни может столкнуться восьмилетний ребенок. Мы советуем оставить максимально простое количество предметов; вы же не хотите запутаться при вычислении среднего или медианного значения, пытаясь определить их для 50 точек данных. Если в комнате есть доска, запишите числа. Можете складывать числа, как это делали мы, если вам нужно использовать один тип среднего значения вместо другого.

A.3.2. Объяснение p-значения

Объясните мне, что такое p-значение и как оно применяется?

ПРИМЕР ОТВЕТА

Представьте, что вы подбрасываете монету и в 26 из 50 случаев выпадает орел. Можете ли вы сделать вывод, что монета «правильная», потому что вам не выпало ровно 25 орлов? Нет! Вы понимаете, что здесь играет роль случайность. Но что, если монета выпадет орлом 33 раза? Как определить порог, при котором можно сделать вывод о «неправильности» монеты?

Здесь на помощь приходит *p*-значение — вероятность того, что при истинности нулевой гипотезы мы увидим результат, который будет экстремальным или более экстремальным, чем тот, который мы получили. Нулевая гипотеза — принимаемое по умолчанию предположение, что не существует связи между двумя наблюдаемыми группами, для которых мы пытаемся доказать обратное. В нашем случае нулевая гипотеза состоит в том, что монета «правильная».

Поскольку *p*-значение является вероятностью, оно всегда находится в диапазоне от 0 до 1. Оно, по сути, отображает, насколько бы нас удивил результат, окажись нулевая гипотеза истинной. С помощью статистического теста можно рассчитать вероятность, при которой «правильная» монета выпала бы орлом или решкой в 33 случаях или чаще (при этом оба результата столь же экстремальны, как и тот, который мы получили).

Получается, что вероятность, то есть *p*-значение, составляет 0,034. Обычно пороговым значением, позволяющим отвергнуть нулевую гипотезу, считается 0,05. В этом случае мы бы отвергли гипотезу о «правильности» монеты.

Если порог *p*-значения составляет 0,05, то мы соглашаемся, что в тех 5 % случаев, когда нулевая гипотеза верна, мы все равно отвергнем ее. Таким образом, мы получаем ложноположительный результат: показатель отклонения нулевой гипотезы, когда она фактически верна.

ПРИМЕЧАНИЯ

Этот вопрос проверяет ваше понимание *p*-значения и способность излагать его определение по существу. Есть распространенные заблуждения о *p*-значении, например что это вероятность получения ложноположительного результата. В отличие от средних значений, о которых мы говорили выше, эту тему можно понимать неверно. Когда вы будете отвечать на вопрос, мы рекомендуем объяснять его на примере. Дата-сайентисты должны уметь общаться с широким кругом стейкхолдеров, включая тех, которые никогда не слышали о *p*-значениях, а также тех, которые уверены, что понимают его суть, но на самом деле это не так. Вы должны показать не только, что понимаете *p*-значения сами, но и можете объяснить его другим.

A.3.3. Объяснение матрицы ошибок

Что такое матрица ошибок? Для чего она нужна?

ПРИМЕР ОТВЕТА

Матрица ошибок позволяет понять, как прогнозы соотносятся с фактическими результатами для конкретной модели. Матрица представляет собой таблицу с 4 различными комбинациями: истинно положительные, ложноположительные, истинно отрицательные и ложноотрицательные результаты. На основании матрицы ошибок можно рассчитать различные параметры, такие как доля правильных ответов алгоритма (процент, правильно классифицированный как истинно положительное или истинно отрицательное значение) и чувствительность, которую еще называют истинно положительным показателем, а также процент положительных результатов, которые были правильно классифицированы как таковые. Матрица ошибок используется в задачах машинного обучения с учителем, где вы классифицируете или предсказываете результат, например прогнозируете, задержится ли рейс или изображена ли на картинке кошка или собака. Приведу пример с рейсом.

	Фактически опоздало	**Фактически прибыло вовремя**
Спрогнозировано прибытие с опозданием	60	15
Спрогнозировано прибытие вовремя	30	120

В этом случае 60 рейсов действительно опоздали согласно прогнозу, но 30, которые, как предполагалось, должны были прибыть вовремя, на самом деле опоздали. Это означает, что истинно положительное значение равно 60/(60 + 30) = 2/3.

Посмотрев на матрицу ошибок, а не на отдельную метрику, вы можете лучше понять производительность модели. Предположим, что вы просто рассчитали долю правильных ответов алгоритма и обнаружили, что она составляет 97 % точности. Выглядит отлично, но может оказаться, что 97 % рейсов прибывает вовремя. Если бы модель просто прогнозировала, что каждый рейс прибывает вовремя, то ее точность составляла бы 97 %, поскольку все рейсы, прибывшие вовремя, классифицируются правильно, но модель была бы совершенно бесполезной!

ПРИМЕЧАНИЯ

С помощью этого вопроса проверяют, знакомы ли вы с моделями обучения с учителем и знаете ли разные способы оценки производительности моделей. В нашем ответе мы поделились двумя метриками, которые вы можете вычислить по матрице ошибок, с помощью которых вы можете показать, что понимаете, как ее использовать, а также рассказали, когда полезно видеть всю матрицу, а не только одну метрику.

A.3.4. Интерпретация регрессионных моделей

Как бы вы интерпретировали эти два набора выходных данных регрессионной модели на основании входных данных и модели? Эта модель построена на наборе данных из 150 наблюдений ириса трех видов: ирис щетинистый, ирис разноцветный и ирис виргинский. Для каждого цветка записываются длина и ширина чашелистика, а также длина и ширина лепестка. Модель представляет собой линейную регрессию, прогнозирующую длину чашелистика на основе других четырех переменных.

ВХОДНЫЕ ДАННЫЕ В МОДЕЛИ

```
  Sepal.Length  Sepal.Width  Petal.Length  Petal.Width  Species
         <dbl>        <dbl>         <dbl>        <dbl>  <fct>
1          5.1          3.5           1.4          0.2  setosa
2          4.9          3             1.4          0.2  setosa
3          4.7          3.2           1.3          0.2  setosa
4          4.6          3.1           1.5          0.2  setosa
5          5            3.6           1.4          0.2  setosa
```

ВЫЗОВ МОДЕЛИ

```
model <- lm(Sepal.Length ~ ., iris)
```

ВЫВОД 1

```
term               estimate  std.error  statistic  p.value
<chr>                 <dbl>      <dbl>      <dbl>    <dbl>
(Intercept)           2.17      0.280        7.76 1.43e-12
Sepal.Width           0.496     0.0861       5.76 4.87e- 8
Petal.Length          0.829     0.0685      12.1  1.07e-23
Petal.Width          -0.315     0.151       -2.08 3.89e- 2
Speciesversicolor    -0.724     0.240       -3.01 3.06e- 3
Speciesvirginica     -1.02      0.334       -3.07 2.58e- 3
```

ВЫВОД 2

```
variable        value
<chr>           <dbl>
r.squared       0.867
adj.r.squared   0.863
sigma           0.307
statistic         188
p.value      2.67e-61
df                  6
logLik          -32.6
AIC              79.1
BIC               100
deviance         13.6
df.residual       144
```

ПРИМЕР ОТВЕТА

На основании итоговых результатов модели можно сказать, что она очень хорошая. Коэффициент детерминации равен 0,867, то есть предикторы указывают на дисперсию 86,7 % в длине чашелистика. Все значимые предикторы *p*-значения составляют менее 0,05. Я вижу, что чем шире чашелистик и чем длиннее лепесток, тем длиннее чашелистик, тогда как более широкие лепестки фактически соотносятся с более короткими чашелистиками. Мы наблюдаем отрицательные коэффициенты у видов ириса разноцветного и ириса виргинского. Так что мы прогнозируем, что у этих видов длина чашелистиков будет меньше, чем для вида ирис щетинистый.

Допустим, мы выявили новый цветок с шириной чашелистика 1, длиной лепестка 2, шириной лепестка 1, и это была разновидность ириса виргинского. Наша модель прогнозирует, что длина чашелистика будет следующей: 2.17 + .496 * 1 + .829 * 2 – .315 * 1 – 1.02, что составляет около 3. Но прежде, чем использовать эту модель, я бы хотел провести еще некоторый диагностический контроль, чтобы понять, нормально ли распределяются остатки. Также я бы хотел найти тестовую выборку, чтобы увидеть, как модель работает на ней, и убедиться в отсутствии переобучения.

ПРИМЕЧАНИЯ

Интервьюер проверяет сразу несколько тем, и вы можете получить дополнительные баллы в зависимости от того, сколько правильных ответов дадите. Здесь проверяется ваше понимание статистики модели (например, коэффициента детерминации), а также оценок и связанных с ними *p*-значений. Хотя вас не спрашивали об этом напрямую, но в примере ответа мы указали, как бы мы применяли эту модель для прогнозирования длины чашелистика нового цветка. Наконец, мы указали, какую информацию хотели бы получить, прежде чем использовать модель. Вопросы открытого типа — хорошая возможность испытать удачу и дать ответ, который хочет услышать интервьюер, а также дополнительно рассказать о том, что вы считаете уместным в конкретном случае.

Не переусердствуйте и не тратьте 20 минут на один вопрос; покажите, что понимаете как можно больше концепций, а затем двигайтесь дальше.

A.3.5. Что такое бустинг?

Что означает термин «бустинг», когда речь идет об алгоритмах машинного обучения?

ПРИМЕР ОТВЕТА

Бустинг относится к целому классу алгоритмов машинного обучения и предполагает взятие слабой модели и ее многократное применение достаточное количество раз, чтобы она стала сильной. Суть в том, чтобы обучить слабую модель на данных, найти области, в которых модель выдавала ошибки, и лучше обучить вторую модель того же типа, которая взвесит точки данных с ошибками и исправит некоторые из них, которые были у первой модели. Процесс повторяется до тех пор, пока не будет достигнут конкретный предел количества моделей. Далее вы используете все эти модели вместе, чтобы сделать прогноз. При большом наборе моделей вы получите более точный результат, чем при использовании одной. Одной из очень распространенных реализаций бустинга является XGBoost, широко используемый как в R, так и в Python.

ПРИМЕЧАНИЯ

«Бустинг» — достаточно редкий термин, поэтому вполне возможно, что человек с базовым опытом в Data Science может не знать, что именно он означает. Поэтому данный вопрос скорее проверяет более глубокие знания, чем наличие опыта. Вопрос немного теоретический; можно представить, что кто-то давно и успешно использует XGBoost, особо не задумываясь о том, как он работает. Про этот вопрос скорее можно сказать так: «Хорошо, если вы ответите правильно, но если

не получится, это еще не конец света», а не так: «Если вы ответите неправильно, то вряд ли получите работу».

A.3.6. Любимый алгоритм

Какой ваш любимый алгоритм машинного обучения? Ок, объясните мне его.

Йерун Янссенс

ПРИМЕР ОТВЕТА

Мой любимый алгоритм машинного обучения — рекуррентная нейронная сеть. В последнее время я много занимался обработкой с использованием естественного языка, а рекуррентные нейронные сети — отличные модели для быстрой классификации текста.

Вы слышали о линейной регрессии? Нейронная сеть похожа на линейную регрессию, за исключением того, что у вас есть группы линейных регрессий, а выходные данные одной группы являются входными данными для следующей. При связывании всех этих линейных регрессий вместе в слои моделей прогнозы получаются гораздо точнее.

Рекуррентная нейронная сеть — это частный случай нейронной сети, которая настроена на данные, выстраивающиеся в последовательности. При обработке блока текста на естественном языке часть выходных данных последовательности слов является входными данными для модели следующих слов.

ПРИМЕЧАНИЯ

Этот вопрос — один из многих, которые могут задать во время интервью. Он нужен для того, чтобы выяснить, умеете ли вы объяснить сложные вещи простым языком. Важно, не то, какой алгоритм вы выберете, а навык четко объяснить, как он работает. Тем не менее этот вопрос — прекрасная возможность рассказать о предыдущей интересной работе и описать соответствующий алгоритм, которым вы пользовались.

A.3.7. Набор данных для обучения и тестовые данные

Что такое обучающие и тестовые данные? Какова ваша общая стратегия создания этих датасетов?

ПРИМЕР ОТВЕТА

Обучающие данные — это те, которые используются для обучения модели МО. Тестовые данные — те, которые не используются при обучении; они нужны,

чтобы проверить, насколько хорошо работает модель. Эти датасеты должны быть разделены, потому что если данные используются для обучения, то модель может узнать правильный результат для данных и научиться подгонять под него ответ. Есть много способов разделения данных для обучения и для тестирования. Обычно я беру небольшую случайную выборку, например 10 %, в начале анализа и использую ее в качестве тестовых данных для всех моделей, тогда как остальные 90 % — это данные для обучения. Когда я нахожу понравившуюся модель, которая работает достаточно хорошо, я переобучаю ее на всех данных (как для обучения, так и для тестирования), чтобы получить наиболее точную модель для развертывания в производство.

ПРИМЕЧАНИЯ

Очень важно хорошо объяснить разницу между обучающими и тестовыми данными, потому что понимание этого лежит в основе создания модели МО. При этом есть множество рабочих стратегий разделения данных. Например, помимо случайной выборки можно использовать перекрестную проверку, чтобы избежать смещения модели при ее обучении на большем количестве данных. Если вы можете логически обосновать выбор определенного метода, все должно быть хорошо.

A.3.8. Отбор признаков

Как бы вы выполняли отбор признаков, если бы у вас было 1000 ковариантов и вам пришлось бы сократить их до 20?

Алекс Хейс

ПРИМЕР ОТВЕТА

Это можно сделать несколькими способами. Одно из возможных решений в случае прогнозирования — использовать регрессию лассо. Регрессия лассо — это особый тип линейной регрессии, которая применяет штраф к увеличению значения коэффициентов. Увеличивая штрафное слагаемое в регрессии, вы можете уменьшать количество коэффициентов модели до тех пор, пока она не будет использовать только 20 наиболее важных ковариантов. Таким образом, модель выбирает, какие коэффициенты должны быть в модели. Хотя регрессия лассо имеет более низкую оценку точности, чем линейная регрессия со всеми ковариантами в обучающих данных, ее преимущество в том, что она использует только небольшое их количество и может лучше работать с тестовыми данными, поскольку лассо снижает вероятность переобучения. Вы также можете использовать методы уменьшения размерности, такие как анализ главных компонентов (principal component analysis, PCA), чтобы уменьшить размерность с 1000 до 20. При использовании метода

лассо будет выбрано 20 функций из 1000 существующих. При анализе главных компонентов будет создано 20 новых функций, которые попытаются собрать как можно больше данных из 1000.

ПРИМЕЧАНИЯ

Есть много возможных решений этой задачи. Одно из них — попытаться использовать ступенчатую функцию для удаления ковариантов до 20. Можно также взять множество выборок датасетов из 20 функций и выбрать наиболее подходящий набор. Задавая этот вопрос, интервьюер проверяет не вашу способность найти правильное решение, а скорее то, сможете ли вы при необходимости в принципе найти его. Так пытаются убедиться, что вы не застрянете над задачей и работа не встанет. Сможете ли вы придумать какие-нибудь варианты, которые стоит попробовать? Если да, то отлично; вы справитесь. В противном случае у вас могут возникнуть трудности, если придется работать самостоятельно.

Если вы предложите несколько вариантов, будьте готовы объяснить, в каких случаях будете использовать каждый из них. Этот вопрос проверяет, понимаете ли вы предложенные методы или просто выбрали их, потому что кто-то о них сказал. В этом случае вы можете ответить, что выбор зависит от целей, для которых проводится анализ: лассо легко интерпретируется, но анализ главных компонентов допускает больше вариативности.

A.3.9. Развертывание новой модели

Вы разработали новую модель, которая работает лучше, чем старая, находящаяся сейчас в производстве. Как вы определите, стоит ли сменить модель? Как вы это сделаете?

Эмили Спан

ПРИМЕР ОТВЕТА

Для меня ответ зависит от нескольких факторов, связанных со средой. Во-первых, по какому показателю новая модель лучше? Если допустить, что это общая точность, я бы проверил, так ли хороша модель, что ее можно запускать в производство вместо старой, и стоит ли оно того. Если точность будет выше всего на долю процента, то, возможно, это не стоит усилий, поскольку эффект будет незначительным. Во-вторых, есть ли риск нарушения работы текущей модели? Если бы модель развертывалась с использованием хорошо обслуживаемого конвейера с четким протоколированием и тестированием, я бы, вероятно, поменял ее, но если это делалось вручную, перемещением в продукционную систему, да еще и человеком, который уже уволился, я бы, наверное, не стал ничего менять.

В-третьих, есть ли способ сначала провести A/B-тестирование модели? В идеале я хотел бы, чтобы старая и новая модели работали параллельно: так можно проверить, есть ли у новой модели какие-либо сложности или пограничные случаи, которые она пропустила. Ни одна тестовая система не может охватить все нюансы производства, поэтому возможность запустить модель сначала для выбранной группы клиентов или входных данных была бы идеальной.

ПРИМЕЧАНИЯ

Развертывание модели часто является трудоемким и рискованным делом для компании. Этот вопрос определяет, понимаете ли вы, что это такое, и как вы подойдете к решению вопроса. Младший дата-сайентист или инженер по машинному обучению может решить, что правильный выбор — как можно быстрее развернуть наиболее точную модель, но нужно при этом учитывать определенные риски. Если у вас есть какой-либо опыт, на который вы можете опираться (например, вы сталкивались с неудачным развертыванием модели), то сейчас идеальный момент, чтобы рассказать о нем. Не страшно, если такого опыта нет; просто попытайтесь описать, что, по вашему мнению, может пойти не так.

A.3.10. Поведение модели

Какими параметрами вы бы руководствовались для оценки разработанной вами модели с точки зрения конечного пользователя? Какие ошибки вы бы считали допустимыми?

Тереза Иофчиу (Tereza Iofciu)
и Бертил Хатт (Bertil Hatt)

ПРИМЕР ОТВЕТА

Стандартные показатели модели, такие как коэффициент детерминации или точность, могут не охватывать необходимые показатели конечного пользователя или бизнеса. Модель классификации может быть верной в 99 % случаев, но в 1 % она будет выдавать неправильный результат, являющийся настолько важным параметром, что от модели откажутся.

Я считаю, что лучший способ оценить модель — это попробовать провести с ней эксперимент. Например, если бы я создавал модель кластеризации для сегментирования клиентов, то передал бы кластеры маркетинговому отделу и попросил бы их провести тестовый запуск кастомизированного маркетинга для выборочной группы клиентов из разных сегментов. Я бы сравнил, насколько улучшились показатели маркетингового отдела при наличии сегментации и без нее, и если бы наблюдалось значимое улучшение, то модель можно было

бы считать успешной. Этот подход полностью отличается от метода, когда мы рассматриваем показатели самой модели, например насколько эффективно она выполняет сегментацию, потому что они оценивают непосредственно только модель. Здесь я фактически анализирую производительность модели в сравнении с производительностью при ее отсутствии.

Недостатком эксперимента с моделью является сложность организации эксперимента. Иногда невозможно разделить клиентов на тех, к кому будет применяться модель, и на остальных. В других случаях эффект от модели настолько незначителен, что он не отражается ни в одном KPI, который легко посчитать. Но несмотря на эти недостатки, если у вас есть возможность провести эксперимент, это почти всегда отличный вариант.

ПРИМЕЧАНИЯ

Это сложный вопрос, потому что он очень общий, но для ответа на него нужно углубиться в детали. Ответы могут быть абсолютно разные в зависимости от того, о чем вы говорите — о моделях прогнозирования или моделях без учителя; ответы также зависят от того, с каким отделом вы работаете — маркетинговым или операционным. Нужно много говорить о том, что статистические выводы — это не то же самое, что показатели, которые интересны для бизнеса; младшие датасайентисты могут излишне сосредоточиться на статистике и не учесть бизнес-показатели. Но выбор стиля ответа остается за вами. Как и в других случаях, если вы можете привести примеры из личного опыта, расскажите о них.

A.3.11. Экспериментальный проект

(Вопрос, ответ и комментарий от Райана Уильямса)
Вы разрабатываете приложение и хотите определить, будет ли новый формат лучше нынешнего. Как бы вы структурировали тест, чтобы выбрать наиболее подходящий формат приложения?

ПРИМЕР ОТВЕТА

Есть много разных способов ответить на подобные специальные вопросы, но в A/B-тестировании обычно применяется такая последовательность:

1. Определите, что вы подразумеваете под термином «лучше», выбрав показатели, которые хотите усовершенствовать: активные пользователи, клики, общее впечатление и так далее.
2. Выберите нулевую гипотезу на основе успешной метрики, например: «Количество кликов будет одинаковым для всех групп». Используйте эту гипотезу, чтобы выполнить расчет статистической мощности, который

покажет, сколько времени понадобится для запуска теста для выявления изменения определенного размера.

3. Случайно разделите совокупность пользователей приложения на группы и предоставьте каждой из них отдельную версию приложения.
4. После того как тест проработает в течение времени, определенного на втором этапе, оцените с помощью соответствующего теста наличие статистически значимой разницы между двумя группами (например, можно применить критерий Стьюдента).

ПРИМЕЧАНИЯ

Подобные вопросы часто задают дата-сайентистам в командах, активно участвующих в мониторинге соцмедиа, разработке приложений, веб-разработке и так далее. Интервьюер обычно просто хочет убедиться, что вы понимаете цель и общие принципы A/B-тестирования, особенно если вы джуниор. Вместо того чтобы вдаваться в специфику статистического тестирования (например, рассказывать, когда следует использовать критерий хи-квадрат, а не критерий Стьюдента), мы рекомендуем продемонстрировать, что вы сможете спланировать эксперимент и определить причинно-следственную связь.

A.3.12. Недостатки экспериментального проекта

(Вопрос, ответ и комментарий от Райана Уильямса)
Предположим, вы провели A/B-тестирование, чтобы выбрать наиболее подходящий формат приложения. В каком случае не следует внедрять новый формат, несмотря на статистически значимое улучшение тестируемого показателя?

ПРИМЕР ОТВЕТА

Этого не следует делать, если вы заметите, что он отрицательно влияет на другие важные показатели (метрики guardrail («ограждение») a. k. a. «не навреди»). Например, тестируется метрика кликов, и хотя вы видите, что посетители, использующие новый формат, действительно переходят по ссылкам чаще, также заметно, что загрузка страниц в приложении занимает больше времени. В этом случае снижение производительности приложения может не стоить увеличения количества переходов, потому что ухудшение качества работы приложения может не понравиться пользователям.

ПРИМЕЧАНИЯ

Это вопрос открытого типа. Интервьюер хочет выяснить, понимаете ли вы, что выявление низкого *p*-значения не всегда является достаточной причиной для того,

чтобы считать эксперимент успешным. Компания рискует, когда в действующий продукт (например, в приложение или веб-сайт) вносятся изменения, а один статистический тест обычно не включает всю информацию, необходимую для принятия правильного решения. Отвечая на этот вопрос, можно также рассказать о соотношении между незначительным улучшением и стоимостью/рисками внесения изменений или о смещениях в выборке/методологии разбиения.

A.3.13. Смещение в выборочных данных

(Вопрос, ответ и комментарий от Райана Уильямса)
О каких типах смещений следует помнить при использовании выборочных данных? Как определить смещение в выборке?

ПРИМЕР ОТВЕТА

На выборочные данные могут влиять многие типы смещений. Одним из наиболее распространенных в практике анализа данных является систематическая ошибка отбора (неправильный отбор выборки). Она может возникать при таких сценариях, как выбор случайной группы клиентов из таблицы уровня транзакции, и перепредставляет клиентов с несколькими транзакциями. К другим типам распространенных смещений относятся систематическая ошибка выжившего (выборка чрезмерно представляет группу, которая прошла некий предварительный отбор) и смещение выборки добровольного опроса (выборка чрезмерно представляет группу, которая с большей вероятностью добровольно предоставила информацию о себе).

Есть статистические методы, с помощью которых можно выявлять смещения выборки, например сравнение среднего значения выборки с известным или ожидаемым средним значением для генеральной совокупности. Чтобы выявить смещения, к процессу получения выборки следует подходить рационально, задавая себе такой вопрос: «Может ли наш способ выбора группы каким-либо образом повлиять на результат, сделав его отличным от генеральной совокупности, которая нас интересует?»

ПРИМЕЧАНИЯ

Этот вопрос задается, чтобы проверить ваше понимание определенных ограничений при работе с данными, а также умение делать выводы. В данном случае важно не столько знать конкретные термины вроде *систематической ошибки отбора* или *ошибки выжившего*, сколько понимать, что данные могут быть ограничены или вводить в заблуждение. Интервьюер хочет убедиться, что вы разбираетесь в нюансах работы с реальными данными, — все они в той или иной степени со

смещениями — и весь беспорядок возникает в результате их использования. Например, использование данных необязательного опроса определенно имеет смещение выборки добровольного опроса. Это не означает, что данные непригодны для использования, но вы должны знать о смещении, думать о последствиях, которые оно имеет для анализа, и учитывать его в любых своих выводах.

A.4. Поведенческие вопросы

A.4.1. Ваш наиболее значимый проект

Расскажите о вашем наиболее значимом проекте, который оказал существенное влияние на работу компании?

ПРИМЕР ОТВЕТА

На последнем месте работы меня попросили скомпилировать систему веб-аналитики, или A/B-тестирование. Компания была заинтересована в проведении экспериментов; у них был инженер, который мог их реализовать, два маркетолога с идеями и пониманием изменений и менеджер, но им нужен был способ понять результаты экспериментов.

Вначале я анализировал каждый эксперимент отдельно в R. Но я знал, что это не лучший выбор, ведь для того, чтобы команда увидела результаты, я должен был запускать сценарии, так что я дублировал работу для каждого анализа.

Именно поэтому я создал внутреннюю панель индикаторов (дашборд) для мониторинга экспериментов. Она включала не только результаты каждого эксперимента, такие как процент людей, зарегистрировавшихся или подписавшихся в контрольной группе по сравнению с экспериментальной, но также проверки состояния работы, чтобы убедиться, что эксперимент выполняется в штатном режиме и что результатам можно доверять. С помощью этой панели управления любой сотрудник компании может увидеть самые актуальные результаты.

К тому времени, когда я уволился, эта панель индикаторов использовалась для всех экспериментов, которые проводили пять команд. Благодаря работе, которую я проделал с остальными членами экспериментальной команды, почти каждый программный сервис компании сначала тестируется в экспериментальном режиме, чтобы определить, дает ли он какой-либо положительный результат.

ПРИМЕЧАНИЯ

Лучше говорить о DS-проектах, которые вы выполняли для компании, а не о тех, которые не связаны с данными. Если же вы выполняли их только для личных целей или в качестве учебного задания, можете рассказать о другом проекте.

Главное — уделить особое внимание тому, какие результаты проект принес для бизнеса. «Я построил модель с 90 %-ной точностью!» не то, что хочет услышать интервьюер; он хочет понять, как другие применяли созданные вами модель, инструмент или анализ и какую пользу принесла ваша работа.

A.4.2. Неожиданные данные

Расскажите о случае, когда вы обнаружили в данных что-то такое, что вас очень удивило.

ПРИМЕР ОТВЕТА

Ранее я работал в компании, которая зарабатывала на подписках. Я работал там над экспериментами, вначале рассчитывая коэффициент подписки как процент людей, которые вошли и подписались позже. На первый взгляд неплохо, но оказалось, что у людей появлялись подписки с датой из будущего!

Поговорив с дата-сайентистом, который собирал эти данные, я обнаружил, что подписки с будущей датой были приостановленными. Возьмем пользователя с ежемесячной подпиской, активированной в сентябре. Чтобы не платить за два месяца и не терять доступ к контенту, пользователь мог приостановить подписку, а не продлевать или отказываться от нее в октябре, а затем возобновить ее в декабре. В этом случае для пользователя в таблице подписок будут две строки: одна для подписки с сентября по октябрь, а другая для подписки с декабря.

В моем случае мне не стоило учитывать подписки, которые активируются в будущем, потому что они были бы возобновлены; мне нужны были только те подписки, которые кто-то активно выбирал!

Я усвоил два урока: первый — никогда не следует делать предположений относительно данных; второй — может потребоваться настройка источника данных в соответствии со своими потребностями. Я предполагал, что подписка с будущей датой невозможна, поэтому не проверял этот параметр. Когда я осознал, в чем причина проблемы, то не стал перезаписывать исходные данные, потому что другим людям все еще нужно было знать о подписках, которые должны будут активироваться в будущем. Вместо этого я сделал собственную таблицу, в которой учитывались только новые подписки.

ПРИМЕЧАНИЯ

В этом ответе мы привели пример, в котором были удивлены тем, что по сути является вопросом качества данных для нашего конкретного сценария. Вы же можете рассказать о ситуации, когда вас подвела интуиция и результат не совпал с вашими ожиданиями, например вы могли проводить исследовательский

анализ данных для DS-подраздела на Reddit, думая при этом, что количество слов в постах будет положительно коррелировать с количеством комментариев, но оказалось, что корреляция отрицательная. Убедитесь, что вы объяснили, почему придерживались своего первоначального предположения.

Этот вопрос проверяет, обдумываете ли вы данные, прежде чем просто погрузиться в работу. Он также проверяет, что вы не просто пытаетесь подтвердить свою первоначальную гипотезу, но и допускаете возможность того, что она не будет соответствовать фактически полученным результатам, и вы готовы это принять.

A.4.3. Размышления о предыдущей работе

Что вы больше всего хотели изменить на предыдущей работе, но не могли?
Бертил Хатт (Bertil Hatt)

ПРИМЕР ОТВЕТА

Я заметил, что на последнем месте работы были реальные проблемы с общением. Руководство постоянно просило сотрудников быть более открытыми и высказываться о потенциальных проблемах, но этого никто не делал. Я предполагаю, что причина кроется в том, что сами руководители не были открыты; они постоянно говорили нам, что все идет отлично, хотя мы знали, что это не так.

Больше всего мне хотелось, чтобы руководство было искренним по отношению к нам, и именно это мне хотелось изменить. Если бы с нами открыто говорили о трудностях и рабочих проблемах, то джуниорам было бы легче открыться в ответ, что улучшило бы рабочую атмосферу.

ПРИМЕЧАНИЯ

Это *непростой* вопрос. Вам нужно показать, что вы достаточно хорошо понимали предыдущую рабочую среду, чтобы предложить ее улучшение, но в то же время вы должны преподнести это так, чтобы создать впечатление, будто у вас были хорошие взаимоотношения с предыдущим работодателем.

Вы можете перечислить множество различных вариантов изменений: технические, командные, продуктовые. Чем более значимое изменение вы предложите, тем лучше (не стоит говорить: «Я бы хотел, чтобы у нас была бесплатная газировка»). Также было бы здорово, если бы вы могли объяснить, почему этого изменения не произошло («Я бы хотел, чтобы мы использовали современный язык вроде R или Python, но поскольку мы поддерживали устаревшие продукты, то приходилось пользоваться SAS»). Если вы можете объяснить, почему желаемого изменения не произошло, это покажет, что вы задумались об ограничениях рабочего места.

Не говорите о предыдущем работодателе оскорбительно («Просто невозможно поверить, что он был до того глуп, что использовал FORTRAN!»). Не стоит созда-

вать себе амплуа сотрудника, который, уволившись, будет оскорблять предыдущего работодателя. Проявляйте уважение, даже если работодатель не полностью соответствует вашим ожиданиям.

A.4.4. Специалист, занимающий более высокую должность, делает ошибку на основании данных

Что бы вы сделали, если бы ваши расчеты или результаты противоречили результатам, полученным ранее одним из руководителей? Вы бы попытались доказать свою правоту? Если да, то как?

Хлинур Халлгримссон (Hlynur Hallgrímsson) и Хизер Нолис

ПРИМЕР ОТВЕТА

Во-первых, я бы спросил себя, достаточно ли важен этот результат, чтобы поднимать эту тему в принципе. Если бы разница составляла небольшой процент и наша команда все равно бы приняла аналогичное решение с новыми результатами или если бы предыдущие результаты никогда ни для чего не использовались, я не обратил бы внимания на расхождения.

Если нет, я бы попытался понять мотивы и цели другого человека. Предположим, что этот человек — вице-президент по продажам и он провел анализ, показывающий, что каждый нанятый им продавец приносил компании доход от продаж в более чем две свои зарплаты. Затем этот анализ использовался бы в качестве обоснования для найма еще пяти человек. Если я покажу, что каждый продавец на самом деле приносит компании доход меньше собственной зарплаты, это может поставить под угрозу весь отдел продаж. Люди будут во многом зависеть от вашего результата, поэтому важно соблюдать осторожность.

В этом случае я бы назначил совещание. Понимая ситуацию, я мог бы предсказать ответную реакцию. Если результаты противоречат друг другу из-за ошибки в анализе вице-президента или если результат имеет фундаментальное значение для бизнеса, я бы ожидал, что человек будет отстаивать свои выводы и попытается найти недостатки в моем анализе. Поэтому я бы подготовился морально и особо тщательно перепроверил бы свои результаты. Я бы попытался найти решение, которое не навредит репутации вице-президента и даст ему возможность изменить стратегию и направить бизнес в правильное русло.

Худший сценарий — это если меня не станут слушать и будут предлагать веские доводы, по которым новые результаты неверны, но я все равно считаю, что они крайне важны для бизнеса. В такой ситуации я бы поговорил со своим руководителем, чтобы он помог разработать стратегию, с помощью которой можно было бы поделиться новыми результатами и принять соответствующие меры. К сожалению, иногда люди не соглашаются с новым анализом, поэтому

необходимо сместить фокус и вместо того, чтобы убеждать человека в своей правоте, нужно найти способ достигнуть своих целей и ограничить последствия неправильного анализа.

ПРИМЕЧАНИЯ

Задавая этот вопрос, интервьюер пытается понять, как вы справлялись с конфликтами, если они связаны с кем-то вышестоящим. На этот вопрос существует множество ответов, но некоторые из них давать определенно не стоит. Например, лучше не говорить: «Я бы отправил имейл всем сотрудникам компании, чтобы все они узнали об ошибке» или «Я всегда буду поступать так независимо от ситуации, потому что единственное, что имеет значение, — это данные». Научные сотрудники могут испытывать особые трудности с конфликтами в бизнесе; в академических кругах такого рода переговоры могут перерастать в состязание на тему «Кто найдет больше всего недостатков в анализе и опровергнет аргументы». В индустрии вы, наоборот, должны уметь делиться своей точкой зрения и помогать бизнесу принимать правильные решения, учитывая другие факторы и понимая нюансы различных ситуаций. Интервьюер хочет понять, есть ли у вас опыт успешного решения разногласий, поэтому постарайтесь его продемонстрировать как можно убедительнее.

A.4.5. Разногласия с членами команды

Расскажите о ситуации, когда вы были не согласны с членами команды. С чем это было связано и над чем вы работали?

ПРИМЕР ОТВЕТА

Однажды мы с продакт-менеджером работали над экспериментом, установив для него продолжительность в две недели на основе расчета статистической мощности. Через четыре дня продакт-менеджер решил досрочно завершить эксперимент и полностью запустить продукт, потому что *p*-значение по основной метрике удачного завершения составило 0,04. Но я знал, что такой результат может быть связан с постоянными просмотрами работы эксперимента: если вы ежедневно проверяете результаты, чтобы посмотреть, опускается ли *p*-значение ниже 0,05, и прекращаете эксперимент, если это происходит, то тем самым вы значительно увеличиваете долю ложноположительных заключений. Я также знал, что продакт-менеджер был очень заинтересован в успешном исходе эксперимента: одним из основных показателей, по которым его оценивали, был дополнительный доход, полученный в результате успешных экспериментов.

В этом случае я решил убедиться, что знаю, откуда человек пришел и почему он задает именно такие вопросы. Я напомнил ему о нашей общей цели:

мы должны сделать компанию максимально успешной. Я привел ему простой пример из веб-комикса *xkcd*, чтобы объяснить, почему преждевременное прекращение эксперимента может стать проблемой: если вы захотите проверить, вызывают ли драже 20 разных цветов прыщи, то даже если знать, что между этими факторами нет никакой связи, все равно есть вероятность, что статистический тест ее «обнаружит». (Вот ссылка на комикс: https://xkcd.com/882/.) Точно так же мы ловили статистических призраков и обманывали себя, думая, что получили положительный результат, хотя на самом деле это не так. В конце концов продакт-менеджер согласился продолжить эксперимент в течение запланированных двух недель.

Эта ситуация заставила меня больше задуматься о том, как я могу улучшить инструмент для экспериментов, чтобы людям было легче делать нужную работу. На одной известной мне экспериментальной платформе есть маленький кружок, который с каждым днем заполняется и через семь дней превращается в галочку. Благодаря этой нехитрой уловке люди проводили эксперименты как минимум в течение недели, что является наиболее эффективной практикой.

ПРИМЕЧАНИЯ

В этом ответе используется модель СОПР (ситуация, отрезки, подход, результат) — это классическая структура для ответов на вопросы поведенческого интервью, поскольку она обеспечивает последовательность ответа, которую легко соблюдать. Продумывая подходящий пример для этого вопроса, вы должны найти ситуацию с положительным исходом; она не должна завершаться фразой: «И с тех пор мы больше никогда не разговаривали» или «Я его уволил». Разногласия должны быть связаны с рабочими вопросами. Не стоит говорить: «Мы не пришли к единому мнению о том, как загружать офисную посудомоечную машину». Интервьюеры хотят понять, можете ли вы с пониманием относиться к оппоненту, не оговаривая и не обвиняя его.

A.4.6. Сложные задачи

Что вы делаете, если не знаете, как решить задачу, связанную с анализом данных?

ПРИМЕР ОТВЕТА

Что касается написания кода, то меня выручает Гугл. Часто, если я гуглю «сообщение об ошибке» или что-то вроде «как выполнить латентное размещение Дирихле в R?», первые страницы поиска обычно выводят ответы в Stack Overflow. Если я знаю, какую функцию или пакет хочу использовать, но не совсем уверен, как она работает, то я проверю это в документации.

Но иногда я не знаю, как подойти к решению задачи. В таких случаях я обычно начинаю разбирать ее на составляющие, порой записываю различные компоненты на доске. Это помогает мне разобраться в основных вопросах, которые, я, возможно, знаю, как решать, даже если изначально вся задача в целом казалась сложной.

У меня есть правило, согласно которому я обычно трачу 15 или 30 минут на попытку самостоятельного решения, и если за это время у меня не получилось продвинуться вперед, то я обращаюсь за помощью к другому дата-сайентисту в компании. Я обязан сначала попытаться найти решение самостоятельно, но в то же время не должен зацикливаться на чем-то целый день, если коллега может помочь мне за несколько минут. Когда я обращаюсь за помощью, то объясняю, что мне удалось выяснить самостоятельно, и привожу примеры, чтобы другому человеку было легче понять суть проблемы (вместо того, чтобы отправлять ему сотни строк кода для анализа).

ПРИМЕЧАНИЯ

Data Science — это область, в которой вы будете постоянно учиться новому и сталкиваться с невиданными прежде задачами, поэтому важно разработать несколько стратегий, благодаря которым вы сможете выходить из сложных ситуаций. Этот вопрос проверяет, есть ли у вас в запасе стратегии для реальных ситуаций, когда перед вами нет вариантов с ответами, одноклассников и профессора, которые могли бы помочь. Возможно, вам придется адаптировать этот ответ под конкретную компанию, в которой вы проходите интервью. Если вы скажете, что ваша основная стратегия — это обращаться за помощью к коллегам, занимающимся данными, и при этом вы претендуете на должность первого дата-сайентиста в компании, такой ответ станет звоночком для интервьюера.

A.5. Вопросы на логику и находчивость

A.5.1. Оценка

Определите, сколько одноразовых бутылочек шампуня используется всеми отелями в Соединенных Штатах в год.

ПРИМЕР ОТВЕТА

Я оцениваю количество бутылочек по следующей формуле:

```
количество отелей в США × среднее количество номеров в отеле × 1 бутылочка
шампуня на каждый занятый номер за ночь × средняя загрузка гостиничных номеров ×
365 дней в году = количество бутылочек шампуня в год
```

Затем я оцениваю значения в формуле:

- *Количество отелей в США*. Если предположить, что на каждые 5000 человек в стране приходится один отель, а в стране проживает около 300 миллионов человек, то всего получаем 60 000 отелей.
- *Количество номеров в отеле*. Думаю, что 50 — вполне достойное число для среднего количества номеров в отеле, если брать за основу те, в которых я останавливался.
- *Средняя загрузка номеров*. Поскольку отели должны приносить прибыль, я предполагаю, что вероятность того, что номер будет занят каждые сутки, составляет 80 %.

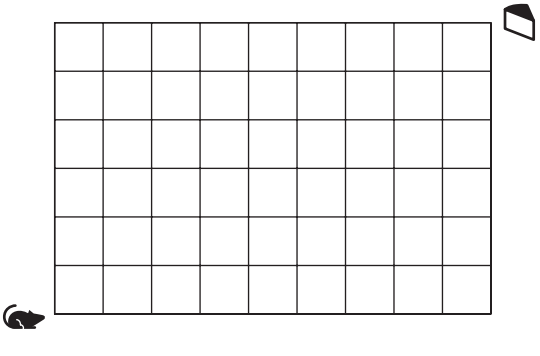
Таким образом, получаем формулу: `60,000 * 50 * 1 * 0.8 * 365 = 876 миллионов бутылочек`.

ПРИМЕЧАНИЯ

Решение этого вопроса заключается в составлении формулы для числа, которое нужно определить, и угадывании чисел, которые нужно подставить в формулу. У этого вопроса есть множество вариаций: «Сколько мячиков для настольного тенниса поместится в самолет "Боинг 747"?» или «Сколько всего роялей во Франции?» Интервьюер хочет понять, сможете ли вы составить подходящую формулу, а также, что вы обоснованно подставляете нужные числа в формулу. Имейте в виду, что у вас практически нет шансов получить ответ, близкий к реальному, непосредственно во время интервью (например, мы не знаем, является ли число 50 хорошим предположением для среднего количества номеров в отеле), но это не важно.

Подготовиться к таким вопросам почти невозможно. Единственное, что вы можете сделать, так это попрактиковаться с ходу придумывать формулы и расчетные показатели для импровизированных параметров.

A.5.2. Комбинаторика



Представьте себе сетку как на рисунке выше. В нижнем левом углу сетки находится мышь, в правом верхнем углу — кусок сыра. Мышь может перемещаться только по линиям сетки и никогда не выходит за ее пределы. Сколько всего путей, по которым мышь может добраться до сыра?

ПРИМЕР ОТВЕТА

Чтобы добраться до сыра, мышь должна девять раз переместиться на одно деление по горизонтальной линии в сетке, а затем шесть раз на одно деление по вертикальной (так как размер сетки равен 9 × 6). Назовем движение по горизонтали H, а движение по вертикали — V. Тогда любая строка с 9 H и V является допустимым путем движения от начала до конца. Например, движение прямо вверх, а затем вправо будет представлено как VVVVVVHHHHHHHHH. Существует 15 факториальных (15!) способов расположить 15 отдельных символов, которые называются перестановками, но, поскольку 6 и 9 из них — это одна и та же буква (V и H), мы должны удалить все символы, дублирующие ходы. Мы можем удалить их, посчитав количество всех повторяющихся действий. V продублированы 6! раз (количество способов их расположения), а H дублируются 9! раз. Это означает, что ответ — 15!/(6!)/(9!), или 5005 путей.

ПРИМЕЧАНИЯ

Этот вопрос и вправду непрост. Во-первых, здесь действительно сложно знать правильный ответ. Если вы хоть немного изучали комбинаторику, возможно, вы найдете решение; в противном случае трудно сразу представить задачу в виде размещения путей. Даже если у вас получится, не факт, что вы будете знать, как подсчитать количество решений.

Во-вторых, даже если вы знаете ответ, будет непросто изложить его в краткой и при этом доступной форме. Не все знакомы с такими терминами, как «перестановка в контексте комбинаторики», и если вам придется все это объяснять, то на это уйдет слишком много времени.

В заключение следует сказать, что вопросов по комбинаторике настолько много, что невозможно заранее подготовить ответы на все из них. Отвечая на подобные вопросы, объясните ход ваших мыслей и то, как вы можете подойти к решению задачи. Если интервьюер уделяет слишком много внимания подобным вопросам, это тревожный звоночек для вас.

Жаклин Нолис, Эмили Робинсон

Data Science для карьериста

Перевела на русский А. Попова

Заведующая редакцией	*Ю. Сергиенко*
Ведущий редактор	*А. Юринова*
Литературные редакторы	*К. Захарцева, М. Столярова*
Обложка	*В. Мостипан*
Корректоры	*М. Одинокова, Н. Петрова*
Верстка	*Е. Неволайнен*

Изготовлено в России. Изготовитель: ООО «Прогресс книга».
Место нахождения и фактический адрес: 194044, Россия, г. Санкт-Петербург,
Б. Сампсониевский пр., д. 29А, пом. 52. Тел.: +78127037373.
Дата изготовления: 07.2021. Наименование: книжная продукция.
Срок годности: не ограничен.
Налоговая льгота — общероссийский классификатор продукции ОК 034-2014,
58.11.12 — Книги печатные профессиональные, технические и научные.
Импортер в Беларусь: ООО «ПИТЕР М», 220020, РБ, г. Минск, ул. Тимирязева, д. 121/3, к. 214, тел./факс: 208 80 01.
Подписано в печать 21.06.21. Формат 70x100/16. Бумага офсетная. Усл. п. л. 29,670. Тираж 700. Заказ